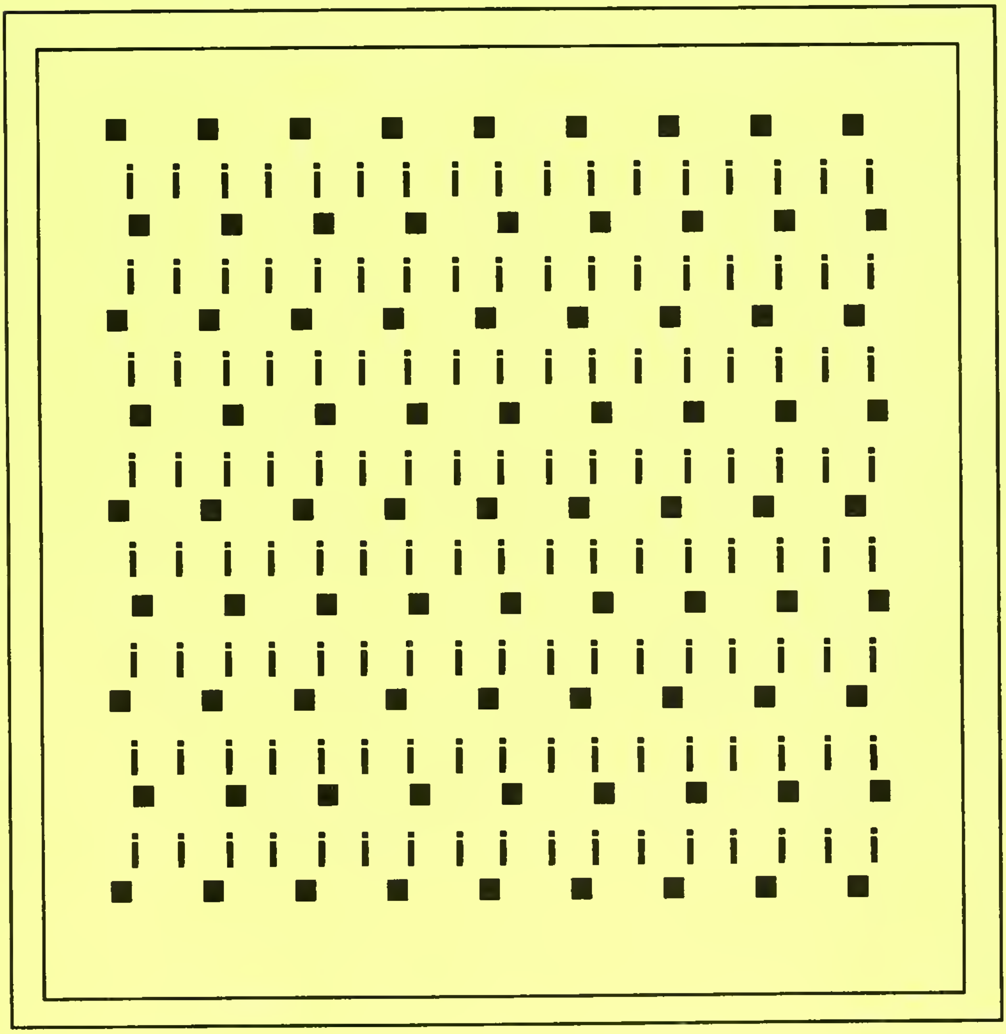
IASSIST QUARTERLY

VOLUME 12 SPRING 1988 NUMBER 1



Printing compliments of The Rand Corporation

IASSIST QUARTERLY

VOLUME 12 NUMBER 1 SPRING 1988

FEATURES

DEPARTMENTS

INTERNATIONAL ASSOCIATION
FOR SOCIAL SCIENCE
INFORMATION SERVICE AND
TECHNOLOGY

• • •

ASSOCIATION
INTERNATIONALE POUR LES
SERVICES ET TECHNIQUES
D'INFORMATION EN SCIENCES
SOCIALES

Editorial Information

The **IASSIST QUARTERLY** represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The **QUARTERLY** reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of **IASSIST.**

Information for Authors

The **QUARTERLY** is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press **Manual of Style** or Kate L. Turabian's **Manual for Writers**. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science 30(2):77-82, March 1979.* If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor:

Walter Piovesan, Research Data Library, W.A.C. Bennett Library, Simon Fraser University
Burnaby, B.C., V5A 1S6 CANADA (01)604/291-4349 E-Mail: USERDLIB@SFU.BITNET

Book reviews should be submitted in duplicate to the Book Review Editor:

Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California
92713 USA. (01)714/856-4978 E-Mail: DTSANG@ORION.CF.UCI.EDU

Key Title: **Newsletter - International Association for Social Science Information Service and Technology**
ISSN - United States: 0739-1137

The Federal Court Data Base:
New research opportunities

by Terence Dungworth
The Rand Corporation
1700 Main Street
Santa Monica, ca 90406-2138

1. Introduction

The Research Division of the Federal Judicial Center has funded the standardization of the District and Circuit Court case records maintained by the Administrative Office of the U.S. Courts. The period covered is FY71 (fiscal year 1971) through FY85.

The product — referred to hereafter as the Federal Court Data Base (FCDB) — was placed in the public domain in the spring of 1986. The intent of the Federal Judicial Center is to update the FCDB at the end of each fiscal year, beginning with FY86. Sections I-V of this document summarize the standardization process and present an overview of the structure and content of the data base. For information concerning further documentation and tape availability, please contact the author.

II. The Reporting System of the Federal District and Circuit Courts

At the end of FY85, the Federal District Court system consisted of 95 districts, organized in 11 geographically structured circuits, plus a single circuit for Washington, D.C. Districts may have one or more offices, with the result that in some districts, cases may be filed and heard in different locations. A list of Circuits and Districts is presented in Table II-1.

All districts process the criminal and civil cases that fall within the jurisdiction of the federal court system. Circuit courts handle appeals from district court dispositions, and some original proceedings.

Certain information on every case filed is reported by each district and circuit to the Administrative Office of the US Courts (AO hereafter) in Washington, D.C. From this is created a central data base containing information on every federal case filed, terminated and appealed in the United States. The reported information on each case is compiled by the AO into a single case record, which, since about 1970, is maintained on magnetic tapes in machine readable format.

Records are grouped by general type — civil, criminal or appeals. Appeals may be from civil or criminal adjudications. At the end of each fiscal year, the AO produces three master tapes containing the records for cases terminated during that year, pending at year-end. Archive tapes are produced for the terminations, and these records are removed from the master tapes for the following year.

The AO also maintains records on cases filed in each fiscal year. These are not incorporated directly into the FCDB, but a comprehensive data base of filings can be constructed from a combination of FCDB terminations and pending files, provided the year(s) for which filings are needed is not earlier than the first year covered by the FCDB (1971).

TABLE 1

FEDERAL CIRCUIT AND DISTRICT COURTS

0090 District of Columbia

First Circuit

0100 Maine
0101 Massachusetts
0102 New Hampshire
0103 Rhode Island
0104 Puerto Rico

Second Circuit

Fifth Circuit

053L Louisiana, Eastern
053N Louisiana, Middle
0536 Louisiana, Western
0537 Mississippi, Northern
0538 Mississippi, Southern
0539 Texas, Northern
0540 Texas, Eastern
0541 Texas, Southern

0205	Connecticut
0206	New York, Northern
0207	New York, Eastern
0208	New York, Southern
0209	New York, Western
0210	Vermont

Tenth Circuit

0311	Delaware
0312	New Jersey
0313	Pennsylvania, Eastern
0314	Pennsylvania, Middle
0315	Pennsylvania, Western
0391	Virgin Islands

Fourth Circuit

0416	Maryland
0417	North Carol., East
0418	North Carol., Middle
0419	North Carol., West
0420	South Carolina
0421	Virginia, Eastern
0422	Virginia, Western
0423	West Virginia, Northern
0424	West Virginia, Southern

0542	Texas, Western

Sixth Circuit

0643	Kentucky, Eastern
0644	Kentucky, Western
0645	Michigan, Eastern
0646	Michigan, Western
0647	Ohio, Northern
0648	Ohio, Southern
0649	Tennessee, Eastern
0650	Tennessee, Middle
0651	Tennessee, Western

Seventh Circuit

0752	Illinois, Northern
0753	Illinois, Central
0754	Illinois, Southern
0755	Indiana, Northern
0756	Indiana, Southern
0757	Wisconsin, Eastern
0758	Wisconsin, Western

Eighth Circuit

0860	Arkansas, Eastern
0861	Arkansas, Western
0862	Iowa, Northern
0863	Iowa, Southern
0864	Minnesota
0865	Missouri, Eastern
0866	Missouri, Western
0867	Nebraska
0868	North Dakota
0869	South Dakota

Ninth Circuit

097X	Alaska

Tenth Circuit

1082	Colorado
1083	Kansas
1084	New Mexico
1085	Oklahoma, Northern
1086	Oklahoma, Southern
1087	Oklahoma, Western
1088	Utah
1089	Wyoming

Eleventh Circuit

1126	Alabama, Northern
1127	Alabama, Middle
1128	Alabama, Southern

0970 Arizona
0971 California, Northern
0972 California, Eastern
0973 California, Central
0974 California, Southern
0975 Hawaii
0976 Idaho
0977 Montana
0978 Nevada
0979 Oregon
0980 Washington, Eastern
0981 Washington, Western
0982 Guam

1129 Florida, Northern
113A Florida, Middle
113C Florida, Southern
113E Georgia, Northern
113G Georgia, Middle
113J Georgia, Southern

III. The Federal Court Data Base Project

The Federal Court Data Base Project (FCDBP hereafter) was funded by the Research Division of the Federal Judicial Center. Its primary objectives were to convert existing machine-readable data into a standardized format.

The need for standardization arose because changes had been made over the years both in variable values that were legitimate and in the formats used to maintain records. In addition, in earlier years, range checking and other validation techniques had not been systematically applied. In consequence, invalid codes could be present in the data for any given year, and inter-year consistency of data structure and content had not been established. This made research time consuming and expensive, and seriously inhibited the utilization of an extremely valuable data resource.

Work on the data was done in two stages. First, the content of all fields for each year was examined, evaluated and cleaned; second, the cleaned data were put into a common format for all years.

The cleaning process was performed on a record-by-record basis. First, values were checked for valid range and format. Frequency distributions of all non-continuous variables were then produced and compared with the list of valid codes for that year. The latter were obtained from AO codebooks and data reporting forms used by the District and Circuit courts. Continuous variables — such as docket numbers, dollar figures, number of months given in a criminal sentence — were evaluated by other approaches.

Invalid codes which were discovered were examined to determine whether or not they could be converted. Frequently this was possible. A simple illustration is a data field that should have been in MMYY format but was reported in YYMM format. This situation was corrected by switching the

YYMM fields to MMYY. Another example is docket numbers which should have had a YYNNNN form (e.g. 800010 would be the tenth case filed in 1980) but had been keyed as 8010 with two trailing blanks. This would be converted to 800010.

If a sound justification for conversion could not be made, variables with invalid codes were assigned missing data values.

The next step was to establish a coding system for each variable that would accommodate all years. Four general principles were adopted as part of this process:

— codes that were effective in 1982 would, where appropriate, be used in place of earlier codes. For instance, the code for the middle district of Florida was changed from 30 to 3A in 1972. FCDB records with Code 30 were, therefore, all changed to 3A.

— all docket numbers were converted to a seven byte field, with the format YYNNNN, where YY is the year of filing, and NNNN is the sequence number of the case within the filing office (not) the filing district. This was due to accommodate the introduction of this format by the AO in 1983. An exception to this rule is cases without a YY indicator in the first two positions of the docket. These were filed before the YY convention was begun. They were right justified in the seven byte field.

— alphanumberic values used by the AO in some ordinal and categorical variables would be replaced by integer values.

— a two byte field would be created for all variables, partly to accommodate future code expansion, and partly to permit the use of negatives (e.g. -8, -9) as missing data codes.

All variables for all years were then converted to standard codes.

The final step was to rewrite the data for each year into a common format. There is one format for each case type (civil, criminal and appeals). Consequently, programs or analytic procedures that work for any one year within case type will also work for any other year.

IV. The FCDB File Structure and Size

The records for any given fiscal year are grouped by district in a single file for each case type, resulting in three terminations files (Civil, Criminal and Appeals) for each year. Counts of the records in each year of terminations covered by the FCDB are presented in Table IV-1. It is possible that some of these records (perhaps one or two in each year) consist entirely of missing data codes. Researchers should accommodate this possibility in their analysis.

Within each year, the organization of records parallels the structure of the court system (see Table 1 above) — i.e. the data are ordered by circuit, district, office within district and docket number within office.

TABLE 2

COUNTS OF DISTRICT COURT CASES AND CIRCUIT COURT APPEALS TERMINATED FY71-FY85 OR PENDING AT START OF FY86

YEAR	CIVIL	CRIMINAL	APPEALS
FY71 Term.	86,564	50,900	12,427
FY72 Term.	95,182	62,500	13,926
FY73 Term.	98,260	59,026	15,092
FY74 Term.	97,634	56,815	15,364
FY75 Term.	104,784	58,911	16,000
FY76 Term.	110,176	59,512	16,358
FY77 Term.	117,151	57,876	17,784
FY78 Term.	125,914	49,727	17,714
FY79 Term.	143,324	44,567	18,928
FY80 Term.	160,482	39,382	20,887
FY81 Term.	177,975	41,017	25,068
FY82 Term.	189,473	43,325	27,987
FY83 Term.	215,356	46,354	28,662
FY84 Term.	243,113	48,325	31,186
FY85 Term.	269,848	50,421	31,387
FY86 Pend.	254,114	32,620	24,761
TOTALS	2,765,328	847,430	333,531

Civil and Criminal terminations for each district are sorted by office and docket number. The docket number alone is not sufficient to uniquely identify a case because different offices within a district may use the same sequence of docket numbers.

For civil cases, there is a single record within the fiscal year of termination. Multiple parties, whether plaintiffs or defendants, are incorporated into this record, with party specific information being taken from the lead party in each group.

For criminal cases, there is a record for each defendant. Office and docket numbers are the same for these records, so they are distinguished by defendant number and name. This creates identification problems when appeals result from multiple defendant criminal cases because the defendant number is not carried forward to the appeals record. Therefore, defendant name is the only way of ascertaining which of the defendants has appealed.

Circuit Court cases, consisting of appeals from District Court decisions and certain original proceedings, are organized by Circuit and Docket Number. Since each circuit uses a single sequence of docket numbers, no additional identification is needed to uniquely specify a case. All appeals records contain the district, office and docket number of the case being appealed, and it is this that can be used to link an appeal to its district court predecessor.

V. Variables Included In The Federal Court Data Base

The original sources of the information included in the FCDB are the forms that Circuit and District Court Clerks forward to the AO in Washington. Separate forms are used for filing and termination, and, during the life of a case, update information may be transmitted as events occur. The AO subjects the information to certain range and validity checks but makes no substantive changes.

The FCDB contains all information reported to the AO in the following categories:

- Filing Location
- Case Identifiers
- Case Type
- Events and Processing
- Adjudication and Disposition

Certain data items created by the AO after reports are received from clerks' offices have been dropped. These are used by the AO for internal identification and control purposes only, and contain no substantive information about the record to which they apply.

The variables in each of these categories have been carefully screened during the cleaning and editing phases of the FCDB Project, and are now represented either by valid codes or missing data (see the codebooks in Appendices C, D and E for details).

Lists of the variables contained within each of the three general case types — appeals, civil and criminal — are presented in Tables 3A, 3B and 3C respectively.

TABLE 3A

INTEGRATED DATA BASE APPEALS CODEBOOK

1. Record Quality Indicator
2. Appeals Court Circuit
3. Appeals Court Docket Number
4. Reopen Code
5. Docket Data (YYMMDD)
6. US as Appellant
7. Appellant Name
8. US as Appellee
9. Appellee Name
10.Appeal from Magistrate's Decision
11.Type of Appeal
12.Nature of Original Proceedings
13.In Forma Pauperis
14.Divisional Office
in Appeals Court Circuit
15.Administrative Agency
16.Jurisdiction
17.Nature of Suit
18.Offense Code
19.No Type
20.District Court Circuit
decision
21.District Court District
22.District Court Office
23.District Court Docket Number
24.Magistrate Indicator
25.Date Filed in District Court
26.Date Notice of Appeal Filed
27.Filing Date Used (YYMM)

28. Transaction Date
29. Transaction Code
30. Disposition
31. By Judicial Action
32. Without Judicial Action
33. Method of Disposition
34. Opinion/Order
35. Original Proceeding
36. Joined Appeal
37. Joined Appeal Docket Number
38. Complete Record Filing Date
39. Last Briefs Filing Date
40. Submission Date
41. Oral Hearing Date
42. Final Judgement Date
43. Case Termination Date
44. Misc. to General Docket
45. Concur./Dissent. Opinion
46. Probable Cause Decision
for Prisoner Petition
47. Who Made Probable Cause
48. Single Judge/Full Panel
49. Counsel Appointed
50. Counsel Continued
51. Counsel Source (if District)
52. Counsel Source (if Circuit)
53. Judge Code #1
54. Judge Code #2

TABLE 3B

INTEGRATED DATA BASE CIVIL CODEBOOK

1.	Record Quality Indicator
2.	Circuit
3.	District
4.	Filing Office
5.	Filing Docket Number
6.	Filing Date (YYMMDD)
7.	Jurisdiction
8.	Nature of Suit
9.	Origin
10.	Residence
11.	Class Action
12.	Termination Judge
13.	Filing Judge
14.	Trial Date (YYMM)
15.	Demand
16.	Filing Magistrate
17.	County
18.	Style
19.	Termination Date (YYMMDD)
20.	Filing Date Used by AO (YYMM)
21.	Disposition
22.	Termination Magistrate
23.	Procedural Progress
24.	Nature of Judgement
25.	Amount Received
26.	Date Judgement Amount was Received (YYMM)
27.	Judgement for
28.	Magistrate Involvement
29.	Other Involvement
30.	Termination Date Used by AO (YYMM)

TABLE 3C

INTEGRATED DATA BASE CRIMINAL CODEBOOK

1. Record Quality Indicator
2. Circuit
3. District
4. Filing Office
5. Filing Docket Number
6. Defendant Number
7. Filing Date (YYMM)
8. Proceeding Code
9. Filing Offense Code
10. DuplicateDefendant
11. Termination Date (YYMM)
12. Transfer Docket Number
13. Transfer Circuit
14. Transfer Defendant Number
15. Transfer District
16. Transfer Office
17. Interval
18. Offense at Termination
19. Major Offense Disposition
20. Counsel
21. Termination Judge
22. Observation Code
23. Sentence Category
24. Statute
25. Sentence Type
26. Prison Term
27. Probation Term
28. Fine
29. Sex
30. Race
31. Birth Year
32. Marital Status
33. Education
34. Prior Record
35. Presentence Investigation
36. Rule 20 Transfer
37. Defendant Name
38. Major Offense Level
39. Termination Offense Level

Issues concerning the bibliograhic citation of machine-readable data files

by Richard Hankinson[1]
Editor, Population Index
21 Prospect Avenue
Princeton, NJ 08540

I would like to take the opportunity provided by this IASSIST meeting not to present a formal, academic paper on a substantive topic, but to set the background for a dialogue between a community that is broadly represented by the IASSIST membership, and a service, represented in this case by Population Index. My understanding of the IASSIST community is that it represents an international group of individuals and institutions concerned with the management, operation, and use of machine–readable data archives. Population Index, in contrast, is an annotated bibliographic journal that covers the world's population and demographic literature. The area of common interest I would like to explore and discuss with you has arisen from a decision that we took some seven years ago to cite not only books, journal articles, and other published materials, but also machine–readable data files (MRDF). This decision was made, primarily, because we felt that many of the demographers we serve are not only computer–oriented but are more used to working with data in

[1] Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Vancouver, British Columbia, Canada on May 19–22, 1987.

machine-readable form than in traditional printed form, and because we felt that, for a variety of reasons, a growing amount of the relevant demographic data — particularly that produced by national statistical offices concerning vital statistics or census -is only available in machine-readable form, and therefore that we are obliged to cite it in such form if our coverage is to be as complete as we wish it to be.

Before coming to specifics, I would like to give you a little background information. Population Index has been providing bibliographic coverage of the population literature for just over 50 years. It is based at Princeton University's Office of Population Research, the first university-based center for demographic study set up in the United States, founded in the 1930s. Population Index consists primarily of an annotated bibliography of the population literature, which includes literature in all languages; the emphasis, however, is on Western and Slavic languages, in which most of the materials intended for general distribution are published. Approximately 3400 citations, complete with abstracts, are produced each year. These are made available in two forms: firstly, in a quarterly journal, with a global circulation of some 4700, sent primarily to members of the international and U.S. professional associations of demographers and population experts and to institutions; and secondly, as a computerized bibliographic data base, POPLINE, available through the MEDLARS system at the National Library of Medicine in Bethesda, Maryland. This is a cooperative venture involving four U.S.-based population centers at Columbia, Johns Hopkins, the University of North Carolina, and Princeton. Population Index is a fully computerized operation working with the University's mainframe computer, an IBM 3081. It is the product of three editorial/bibliographer professionals (one of whom is a data base specialist), with an administrative/clerical/data entry support staff. Funding is provided partly from federal sources (including NICHD and USAID), partly from subscriptions provided to professional associations (the Population Association of America and the International Union for the Scientific Study of Population), and the remainder from paid subscriptions to the quarterly journal.

Tabel 1: Mrdf Citations 1980-1986 by year and country

	1980	1981	1982	1983	1984	1985	1986	sub-total	total
World					1	6	2		9
Australia	3	1	-	3	-	4	10		21
U.S. Census	5	5	6	28	27*	11	17	99	
U.S.NCHS	-	8	3	-	2	-	10	23	141
U.S. Other	2	1	1	1		10	4	19	
Malaysia	-	-	2	-					2
Canada					2				2
France					2				2
Brazil						1			1
Scandinavia							2		2
Israel							6		6
TOTAL	10	15	12	32	34	32	51	141	186

The subject area covered concerns population and demography, which includes such concepts as population size and growth, spatial distribution, mortality, fertility and family planning, nuptiality and the family, migration, historical demography, population characteristics, population policy, population statistics (including censuses, surveys, and vital statistics), and the relationships among demographic factors and socioeconomic development, natural resources, and the environment.

Over the seven-year period since 1980, Population Index has cited 186 MRDFs (see Table 1). Of these, just over 75% have concerned the United States, some 53% from the U.S. Bureau of the Census alone. While it is reasonable to expect a preponderance of U.S. MRDFs in a listing of this kind, the geographic breakdown indicates either a surprising lack of MRDF containing population data from other countries or a failure of the existing information services — including Population Index — to identify such files and provide information on them to the user public.

Apart from the United States, only Australia seems to be adequately represented over the full period. The 15 citations relating to Malaysia, Canada, France, Brazil, Scandinavia, and Israel were largely the result of information originally gathered from the Guide to Resources and Services produced annually by the Inter-University Consortium for Political and Social Research (ICPSR),[2] and it is at least possible that they represent only the tip of the iceberg. We know that the official statistical agencies of many of the European countries are creating MRDFs containing demographic data. If so, where is the information on these products available on a regular basis? We certainly have not been able to provide our users with a steady flow of information on new products through the 1980s. These and other related questions are at the top of our agenda at this time, and we have come to Vancouver to help find some answers. We particularly want to determine to what extent our lack of success in citing such files is our own fault, which we can and will rectify when we know how to do so, and to what extent it is due either to the lack of adequate information from the creators of such files—which is a more difficult situation to change, but may be one where action is needed—or to inadequacies in information services such as those provided by ICPSR. The same questions may well be posed for other parts of the world. Japan is a case in point in the developed world. As for developing countries, the situation is variable according to our limited information. We have extensive information concerning the MRDFs created during the course of the World Fertility Survey and the Contraceptive Prevalence Surveys, primarily during the 1970s.[3] Furthermore, the Dynamic Data Base in Voorburg, the Netherlands, has an on-going program to add new files concerning survey data on fertility.[4] Latin America has several countries generating MRDF with demographic data, thanks in part to efforts by the Centro Latinoamericano de Demografia (CELADE), a U.N. organization in Santiago, Chile, that has worked with countries for many years and built up a data base of census and survey data sets for Latin America. They also have an ongoing information program that publishes information on a regular basis concerning its holdings in this area.[5] It is unlikely that there are extensive demographic MRDFs in Africa at this time, but they probably exist

[2] Inter-University Consortium for Political and Social Research (ICPSR), Guide to Resources and Services, 1984-1985. Ann Arbor, Michigan. 525 pp.

[3] International Statistical Institute (ISI) and International Union for the Scientific Study of Population (IUSSP). Dynamic Data Base. Catalogue of Survey Data Files. 153 pp. Voorburg, Netherlands. January 1987.

[4] Cleland, J.G. "A new service for demographic analysis: the Dynamic Data Base. Population Index (Princeton, N.J.) 52(4), pp. 540-7. Winter 1986.

[5] United Nations. Centro Latinoamericano de Demografia (CELADE). Bulletin of the Data Bank (Boletin del Banco de Datos), Santiago, Chile. No. 11, LC/DEM/G.39, April 1986. 52 pp.

in several Asian countries if we could track down the available information.

The next question I would like to raise concerns what elements are necessary to a complete bibliographic citation of an MRDF. The policy followed by Population Index is based on guidelines suggested by Judith Rowe,[6] which were in turn based on standards developed by the American Library Association's Subcommittee on Rules for Cataloging Machine Readable Data Files, and the work of Sue A. Dodd within the IASSIST framework. They are consistent with the ANSI standard and its application for social science data files as enunciated by Dodd.[7] The elements of a citation we have been trying to include are as follows:

1. Authorship: full name of author(s) or corporate body responsible for the intellectual content of the file (e.g., principal investigator, project director, or sponsoring agency).

2. Title: full title (no acronyms) containing descriptive words or phrases and dates.

3. Subtitle: secondary title to amplify or restrict main title as in serial or multi-part works.

4. General material designator: denotes the generic form or type of material cited, e.g. MRDF.

5. Statement of authorship: indicates the relationship of the work to the person(s) or corporate body named in the author heading or to other significant parties such as principal investigator, sponsor, or even funding agency.

6. Edition: provides users with valuable information on the source, date, or revisions; new editions indicate addition or deletion of data elements, variables, or fields; recoding or restructuring of the file; addition or deletion of logical records (cases, observations, etc.) or, in the case of programs, changes in the programming language.

7. Imprint: consists of the producer statement and the distributor statement.

8. Extent of file: the MRDF equivalent of number of pages is number of logical records. Other physical characteristics such as tape size, recording density, etc., are excluded since distributors normally offer various recording options. This information is subject to change and it does not affect bibliographic identification, which is determined by the content of the file and not by the container.

9. Accompanying material: includes codebooks, reports, manuals, and other associated material in printed or machine-readable form. Such materials may or may not warrant separate citation.

10. Series statement: collective title(s) and item number or other designation within the series.

[6] Rowe, Judith S. "Population Index to cite publicly available machine-readable data files." Population Index (Princeton, N.J.), 45(4), pp. 567-75. Winter, 1979.

[7] Dodd, Sue A. Bibliographic references for numeric social science data files: suggested guidelines. Journal of the American Society for Information Science (Washington, D.C.), 30(2), pp. 77-82. 1979.

In addition, each citation includes an appropriate abstract.

The main problems we have had concerning the preparation of MRDF citations have been as follows:

1. • No. 5. Statement of authorship. Since we do not have a field for this, we have been including this information, the few times we have found it appropriate, in the abstract following the citation.

2. • No. 6. Edition. We have not found it possible to provide the necessary information in this field because we have been unable to obtain it. We suspect that this information is available, but to date when we have asked MRDF distributors or publishers for information about whether the file they are listing is a revision or edition, they have not been able to supply us with the answers. At this stage, we really do not know if there is a problem here or not. How often are the MRDFs of demographic interest revised and updated? Most of the U.S. examples we are aware of, such as the Current Population Survey, can be cited as new files rather than revised files. Certainly, Population Index has not recited any MRDF in the last seven years because we have learned that revised and updated editions are available.

3. • No. 8. Extent of File. We have had problems in this field primarily because we are editors and bibliographers rather than computer experts. We understand that the key facts to include are number of logical records and logical record length, and we add this information in the abstract if we know it. However, the information we receive does not always include this information in a way we can understand it. Furthermore, the ICPSR Bulletin, for example, as far as we can establish does not specify whether it is a question of logical or physical records. The objective should be to provide the necessary information so that a potential acquirer or purchaser of the MRDF being cited knows whether he or she can mount and use the MRDF on the equipment available. The question we have to clarify is what that information needs to be.

4. • No. 10. Series Statement. We don't seem to pick up or include any information in this category.

There is one related item we would like guidance on. It is our custom now to add information for each citation on the source of information (e.g. the ICPSR Bulletin) and on the location of the file in question (e.g. U.S. National Technical Information Service). Would it be useful to add to this information addresses for the location of the agency from which the file could be obtained? Or can we assume that most people communicate with the appropriate government agencies by phone and that these numbers are easily obtained by those who need them?

We are also not sure whether we need to add information about the release status of the data contained in MRDF. Although many such files containing demographic data are publicly available, the data in them is often restricted: in many cases, no data can be released without the specific authority of the government of the country concerned. Should we cite MRDF files with restrictions of this kind -and if so, what information should we provide on such restrictions?

Another issue that has concerned us is the relationship between MRDF and hardcopy or published materials. The general principle we follow is that we will cite both the report and the MRDF if they are complete in themselves. If the hard-copy materials are only codebooks, reports, and manuals that are to be used in conjunction with the MRDF, they are not cited separately but are merely referred to in the abstract following the MRDF citation. In other words, we feel that citing the same material in two different forms, paper and MRDF, is not an unnecessary duplication given the different needs of our users.

In conclusion, we are struggling to expand the service we are offering to the demographic community by preparing citations and abstracts to appropriate MRDF. As this review has indicated, we have had some success but have run into two main obstacles. The first is our lack of information concerning the availability of new MRDFs in this area, and we need to work with the IASSIST community to establish the extent to which the information exists and we are missing it, and the extent to which it does not exist and to which we can encourage those responsible to provide it. The other obstacle is interdisciplinary: with our bibliographic, editorial, and library backgrounds, we have a familiarity and expertise with the printed word. Our efforts to include MRDFs involve a new concept and language that we will be struggling with for some time.□

Incorporating catalogue records for machine-readable data files into the University of Alberta library's DOBIS public catalogue

by Jana M. Lamont[1]
University of Alberta
Computing Services Data Library

Background

The Data Library at the University of Alberta belongs, organizationally, under the Department of Computing Services. The Data Library was started in 1977 as a registry of machine-readable data held on the University of Alberta campus. To maintain this registry, an on-line data base was established containing entries describing registered machine-readable data holdings. This on-line catalogue was created using SPIRES - the Stanford Public Information Retrieval System.

In 1981, the Data Library acquired a valuable collection of data files of public opinion survey data, donated by the Department of Political Science.

In 1982, arrangements were made to transfer responsibility for the University's membership in the Inter-University Consortium for Political and Social Research (ICPSR) from the Department of Political Science to the Data Library. Through this membership, the Data Library has acquired many research studies relating to a wide range of subjects. In addition, the Data Library's collection has been enriched from local deposits by the Population Research Laboratory (the Edmonton Area Surveys), the Department of Geography (Alberta weather data), and the inclusion of the World Data Bank II geographic coordinate files. Currently the Data Library collection contains close to 1800 machine-readable files from 400 research studies.

[1] Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Vancouver, British Columbia, Canada on May 19-22, 1987

The Proposal and Acceptance Stages of the Project

One of the concerns of the Data Library has been to promote the use of the Data Library's collection of data files and other services such as reference and acquisition, help with the analysis of data files, construction of instructional data sets, as well as archiving and dissemination of data files deposited by researchers. Since 1981, Data Library staff have struggled to provide information about the Data Library, and have been quite sure that there are many users who are not aware of the Data Library's resources and services. One of the reasons for this was that access to information on the Data Library's holdings was available only through an on-line catalogue using the SPIRES system on MTS (Michigan Terminal System) at the University of Alberta Computing Services. Access to this catalogue is only available to those with a Computing Services Signon ID, and for this one must pay monthly charges for usage of computing time, storage, etc.

It was therefore necessary to offer an alternative way of accessing the information on the Data Library's holdings. The best way to achieve this was to make the contents of the data file collection accessible through the University Library's on-line public catalogue, which is available to all library users free of charge. The public catalogue is available on the OAS (Office of Administrative Systems) computer through an IBM-generated library integrated system called DOBIS. The Data Librarian and the Data Library Analyst wrote a proposal to incorporate Data Library catalogue records into the DOBIS public catalogue. This proposal was submitted for approval to the management of Computing Services. With the approval of the proposal at that level, a letter was written to the Chief Librarian outlineing the request for incorporation of Data Library holdings records in the on-line public catalogue and stressing the value of such a service to the university community as follows:

> The Data Library's holdings provide a valuable research resource; and as the demand increases for a general level of computer literacy among the students, the Data Library holdings offer a valuable teaching resource. By providing references for Data Library holdings within the University's central catalogue, there will be an increased likelihood that instructors and students will discover and make use of these materials.[2]

The University Library was also invited to nominate a representative to the Data Library Advisory Committee. Through communications with this representative, Data Library staff learned that the request to merge the cataloguing records into the main catalogue had been referred to the Library Steering Committee on Automation and that a recommendation had to be made on general policy on the inclusion of holdings of collections outside the Library system, as this was the first time that the Library had received such a request.

In the meantime, the Data Librarian had met with the Head of Library Systems and the Head of Cataloguing Division and explained the proposal to them. A few months passed before the Data Library was asked to provide a sample catalogue record for a data set to the Cataloguing Division for examination.

About a month after submitting the sample, the Data Librarian was invited to attend a meeting of the Library's Cataloguing Division to discuss the implications of the proposal and to answer

[2] Jana Lamont and Charles Humphrey, "A Request to Incorporate Catalogue Records of the Computing Services Data Library into DOBIS Public Catalogue," Letter to the Chief Librarian, p. 1.

questions about the Data Library and its operations. These questions concerned the following: the Data Library acquisitions per year, funding, type of data collected, cataloguing standards, authority files, call-numbers used by the Data Library, and staff training in the Cataloguing Division. It was agreed that Data Library staff would undergo a short training course in the Library's cataloguing procedures, and would conform to the Author and Subject authority files of the University of Alberta Library. After this meeting, recommendations were made to the Chief Librarian to accept the proposal for incorporating the Data Library's catalogue records into the Library's public catalogue.

The Training and Production Stages of the Project

A month later, Data Library staff began a brief training course in the Cataloguing Division. This training course consisted of three or four sessions on cataloguing according to the "UTLAS MARC Coding Manual for Monographs"; University of Alberta Library receives the bulk of its cataloguing from University of Toronto Library Automation Systems. The Data Library acquired the above noted UTLAS manual, other required cataloguing tools such as the "University of Alberta Library Cataloguing Procedures Manual", and UTLAS manuals such as "Format for Standardized MARC Bibliographic Records". After four sessions in the Cataloguing Division, the Data Librarian was able to start the preparation of Data Library catalogue records for incorporation into the public on-line catalogue.

Since the Computing Services Data Library follows the rules and standards set forth in Chapter 9 of "The Anglo-American Cataloguing Rules, Second Edition" and "An Interpretative Manual for Cataloguing Machine-Readable Data Files" by Sue Dodd, there were no problems either with main entry or with the physical description of machine-readable data files. The Computing Services Data Library does not use the MARC format in its SPIRES on-line catalogue, but the cataloguing fields in the SPIRES data base correspond to MARC fields and therefore there was no difficulty in creating a MARC format for the Data Library's catalogue records on the Library's cataloguing data entry form. MARC format records are generated by a SPIRES output format which prints on Library forms the information for MARC fields from 100 to 830. The information in fixed fields and a few other fields not used by the Data Library but required by the Main Library, such as fixed fields 1000 to 1058 and variable fields 043 (geographical code), 045 (chronological code), and field 090 (the local call-number), are entered manually on the same computer printed form. A second format, also generated by SPIRES, produces a temporary shelf list card. These two forms are then passed on to the Cataloguing Division. Here, authors' names and the Library of Congress subject headings are checked against the Library's author and subject authority files, following which the cataloguing data are passed to UTLAS via telecommunications. The Data Library receives UTLAS generated products, including spine labels and pockets for the printed documentation, and shelf list cards.

In the six months following staff training, approximately 90% of the Data Library's catalogue records were submitted to the Library for incorporation into the public catalogue. These were subsequently passed to the UTLAS data base and then transferred to the DOBIS system at the University of Alberta.

Currently Data Library staff are adding cataloguing data for new acquisitions only; the retroactive conversion of Data Library records from SPIRES to DOBIS has been completed.

It should be noted that, while the training and production stages of this cooperative project took approximately seven months, the approval stage took close to a year. In any case – the results were well worth the effort: a number of recent enquiries to the Data Library have resulted from searches of the DOBIS public catalogue.□

Bibliography

Anglo-American cataloguing rules. Second edition. Chicago: American Library Association, 1979.

Dodd, Sue. Cataloguing machine-readable data files: an interpretive manual. Chicago: American Library Association, 1982.

University of Alberta. Library. Cataloguing Division. Cataloguing manual. Edmonton: the University, 1987.

UTLAS Inc. Format for standardized MARC bibliographic records. Toronto: UTLAS Inc., 1985.

UTLAS Inc. UTLAS MARC coding manual for monographs. Toronto: UTLAS Inc., 1984.

IASSIST 1989 conference in Jerusalem.

A reminder that IASSIST89 will be held in Jerusalem. The information to date with regard to the conference is as follows:

Place : Jerusalem, The Hebrew University Campus, Faculty House
Time : May 15-18, 1989

Program Committee (not final) :

- Yoel Haitovsky, Hebrew University
- Craig McKie, Statistics Canada
- Judith Rowe, Princeton University
- Michal Peleg, Hebrew University

International Arrangements :

Nancy Hafota,
Social Sciences Data Archive
Hebrew University
Mount Scopus, Jerusalem 91905 ISRAEL
BITNET : KGUNH@HUJIVM1

Further details on the conference will be published in the Quarertly when they become available.

POLLing for data

by Tom W. Smith

Each year scores of organizations conduct hundreds of surveys asking thousands of questions. This mass of data is piled on top of thousands of surveys collected over the years and is soon to be buried beneath an even greater mass of future surveys.

Until recently, survey organizations and archives struggled valiantly but hopelessly to master this great data mass. It was usually possible to keep track of the studies or surveys being conducted, but since most surveys have eclectic content (Smith, forthcoming) it was difficult to keep track of the actual data (i.e. the questions) or to know what had or had not been asked. Recent developments in computerized data retrieval have made major advances towards mastering the data mass (Vavra, 1986).

At the Roper Center for Public Opinion Research, University of Connecticut, POLL (Public Opinion Location Library) became available to the public in 1986. It can be accessed from anywhere in the United States or Canada. This SPIRES-based data retrieval system has over 85,000 survey questions stored in its memory. It contains most major American public opinion polls conducted since the early 1970s such as NORC's General Social Survey, the Gallup Poll, the Harris Survey, NBC, CBS/NYT, and so forth. The holdings are up-to-date, adding the latest polling data as soon as they are released and the historical depth of the data base is being rapidly expanded, with Gallup surveys back to the mid-1960s slated for inclusion by the end of 1987.

POLL searches can be conducted by specifying (a) the general topics or subject headings of interest, (b) the words used in the question text, (c) the date of the survey, and/or (d) the organization that collected the data. For example, on June 16, 1987 I used POLL to locate questions dealing with the death penalty. As the example below illustrates (Figure 1), I used the WORD search option to identify 184 questions that probably dealt with attitudes toward the death penalty. I then employed the SUBJECT option to check for inappropriate results by isolating those questions that were not classified as dealing with Crime. I then inspected these items and found that they were relevant enough for inclusion, so I used BACKUP to restore the previous total. Finally, since I already had all the available NORC questions, I excluded NORC items. That left me with 168 questions about capital punishment.

Figure 1

A Search for Death Penalty Questions

```
FIND WORD DEATH AND PENALTY

Result: 172 items

OR WORD EXECUTE

Result: 175 items

OR (WORD CAPITAL AND PUNISHMENT)

Result: 184 items

AND NOT SUBJECT CRIME

Result: 14 items

ITEMS

[After inspection of these 14 items, it was decided to include them in the search.]

BACKUP

Result: 184 items

AND NOT ORGANIZATION NORC

Result: 168 items
```

Having now isolated the questions that I wanted, I was ready to retrieve the data. There are four ways to obtain output from POLL. First, I could have it printed immediately on my terminal or PC printer. Second, I could have it downloaded into a file on my PC (not an option if I were using a dumb terminal). Third, I could have it printed at the Roper Center and mailed to me first class. Finally, I could have it printed at the Roper Center and sent overnight delivery. The option selected depends on the urgency of one's need and the size of the output.

Whatever procedure is selected, the output will appear as a series of question wordings which can be sorted by date, organization, and other elements. As figure 2 demonstrates, the output contains the full text of the question, the marginal distribution of the responses, and appropriate documentation such as the agency that collected the data, the population sampled, the sample size, the mode of administration, the field dates, the source document and its date, and the subjects under which the question is classified.

Figure 2

Sample Output from Death Penalty Search

```
Question:

     Q011 ARE YOU INFAVOR OF THE DEATH PENALTY FOR PERSONS
     CONVICTED OF MURDER?

Responses:

     YES                               49%
     NO                                40
     NO OPINION                        11

Survey Organization:          Gallup Organization
Population:                   NATIONAL ADULT
Population Size:              1558
Interview method:             PERSONAL
Beginning date:               OCT 29, 1971
Ending date:                  NOV 2, 1971
Source Document:              Gallup Poll—AIPO
Date of Source Document:      NOV 2, 1971
Subject:                      DEATH
                              CRIME

FULL QUESTION ID: USGALLUP.839.Q011
```

The POLL system has many applications. Mostly it is used to identify questions related to a particular research interest. In some cases such a general review of public attitudes towards a topic or a consideration of aggregate trends, POLL can supply all the information that a researcher needs. Or similarly, if one is trying to devise a new survey and is looking for baseline items to replicate, POLL can supply the questionnaire designer with the needed information. In other cases POLL will merely open the door and point the way. For example, the capital punishment output cited above not only identifies particular questions on capital punishment, but also locates studies that have particularly rich or interesting sets of questions on the topic. Machine-readable data sets for most of these studies are available from the Roper Center.

POLL can also be used for other purposes besides general, topical searches. As a specialized locator, POLL can be used to find specific questions that the researcher already knows about. One merely has to specify enough words to uniquely identify the target question. As an analytical tool, POLL can be used to determine the frequency of occurrence of items by organizations and/or time. For example, Table 1 shows how many questions dealing with Blacks appeared on surveys from 1980 to 1986.

TABLE 1

Questions About Blacks, 1980–1986

Date	% of all Questions Referring to Blacks	Total Number of Questions
1980	0.8%	8638
1981	2.4%	6456
1982	0.6%	7932
1983	1.1%	8272
1984	1.9%	10154
1985	0.9%	7202
1986	2.8%	5125

The data base accessed by POLL has great depth and wide diversity. The capital punishment search detailed above produced 184 relevant items and there were between 46 and 193 items on Blacks each year from 1980 to 1986. To illustrate the broad range of items covered by the POLLS system, and to show how wide ranging contemporary public opinion polls are, I chose 10 words at random. As shown below, in all but one case questions were located using that word:

Word Searched For:	Number of Questions Found
New Zealand	11
India	110
Costa Rica	4
Blender	2
Submarine	30
Skating	23
Carrot(s)	2
Moon	31
Gall Bladder	2
Blimp/Dirigible	0

Finally, POLL is so user friendly that it could be charged with fraternizing. While demonstrating the system to a colleague and having shown him how it worked for about 10 minutes, I was called away to the phone. On my return, one half hour later, I found that he had not only learned the system, but had been able to complete the searches he was interested in doing. The system documentation (Roper Center, 1987) is also clear and easy to follow.

In sum, POLL is a rich (and expanding) data base of survey questions, has search and retrieval options that allow the identification of relevant questions, and is easy to learn and use. With POLL, we can master the data mass.□

References

1. Roper Center. User manual for POLL: the Public Opinion Location Library. Storrs, CT.: Roper Center for Public Opinion Research, 1987.

2. Smith, Tom W. The art of asking questions: 1936-1984. Public opinion quarterly forthcoming.

3. Vavra, Janet K. Using SPIRES: The ICPSR experience. IASSIST Quarterly, 10(3): 41-50, Fall, 1986.

Index of machine readable data files for women's studies

by Elizabeth Stephenson and Martin B. Pawlocki[1]
University of California, Los Angeles
Institute for Social Science Research

Methods for indexing data files and the terminology to be used have been discussed and pondered by archivists for some time. Some archives have developed indexing strategies with varying degrees of success. For instance, the Data Archive on Adolescent Pregnancy and Pregnancy Prevention has used SAS to produce topic–by–topic matrices of each question within an individual file, and uses this procedure to index a group of related files, as well as an entire collection. Other archives, such as the Roper Center, have produced detailed, question–level indices to data files. Through the use of CDNet, users of ICPSR data can search a variable–level index to relevant files among a set of pre–selected surveys. The Zentralarchiv für empirische Sozialforschung has produced a mechanism for retrieving and analyzing data from the Eurobarometers, using an indexing vocabulary developed by the Standard Study Description format. Still other archives, such as the State Data Program at the University of California, Berkeley, and the Henry A. Murray Research Center at Radcliffe College, have developed detailed methods for describing and assigning terminology to data files, but lack a method for automatically accessing such information.

[1] Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Vancouver, British Columbia, Canada on May 19–22, 1987

Discussion on how researchers can best retrieve information on survey content is also ongoing, focusing primarily on the level of specificity at which relevant details are described. At the most general level, should a single term by used to encompass an entire survey, such as "elections"?, or, should a group of key words be selected to represent the major elements within a study? At a more specific level, should an item-level index be created?, or should it be possible to perform keyword searches of questionnaires? Clearly there are more questions than answers.

This paper describes a project conducted at the Social Science Data Archive at the Institute for Social Science Research, University of California, Los Angeles. The Archive was asked by the National Council for Research on Women to be a test site for the evaluation of a thesaurus of non-sexist terminology. We were to use this thesaurus as a controlled vocabulary with which to index materials in our collection. We decided to use this as an opportunity to address some of the questions on how to index data files. We also intended to produce a directory of data files held at our facility that would be useful in the study of women and women's issues.

The project had two phases. First, we tested the thesaurus and indexed selected studies; then, we developed a set of dBASE III programs that stored the index and produced a printed directory. In the future, these programs will be augmented to permit on-line searching of assigned index terms. The details of file selection, indexing and development of programs will be described below.

Data file selection

When the project began we intended to index the entire collection of data files. But after browsing through the collections of other archives, we realized that we would have to choose those files which would be most useful in gender studies. Our selection criteria focused on (1) studies held locally by the SSDA, (2) studies containing large samples of female respondents, (3) studies that focused on women's issues, (4) studies with content that would be significantly useful in studying women as a population group. We included surveys, public opinion polls, enumerations, and administrative records. We tried to cover all time periods and geography.

Thesaurus structure

The thesaurus is based on other thesauri and subject classification schemes. Terms are included in the thesaurus to reflect a mid-level of specificity. This had an effect on the level of specificity and detail we were able to maintain during the indexing project. Over 3000 terms having "primary significance to women" are included in the thesaurus and these were used as a controlled vocabulary. Scope notes and broader, narrower, and, related terms are given. Terms are arranged alphabetically. The thesaurus is accompanied by four other display structures: (1) broad-narrow display showing up to four links of broader to narrower terms, (2) rotated or "permuted" display in alphabetical order, (3) grouped term display which organizes terms alphabetically in eleven subject groups, (4) delimiters display of terms used to further categorize topics under a single term.

Indexing process

Since we could not assign index terms using the machine-readable data files themselves, we referred to study documentation such as catalog descriptions, codebooks, questionnaires and related publications. The primary source of information was the codebook.

As part of the indexing process, we decided to run a check on inter-indexer consistency. This gave us a way to judge whether or not we were using terms uniformly, and indexing files at a consistent level of specificity. In our case, in keeping with the structure of the thesaurus, we attempted to maintain a mid-level of specificity. Thus, we tried to index groups of content-related questions, rather than focusing on the study as a whole on the one hand, or on each individual question on the other. We found that problems arose over whether to index questions, groups of questions, or entire surveys. The problems varied depending on the type of data being indexed. We found that we used a much more detailed approach when the data were from a public opinion poll and broader when the data were from a general survey of women.

When all studies had been indexed and entered into the dBASE III format, it was necessary to study the resulting set of terms as a whole. In some cases, we collapsed two or more terms into one category, since the terms were so similar that their individual use was unnecessary. In other instances we grouped more specific terms into a broader category to maintain a mid-level of specificity. In general, we avoided making these changes. Often there is only one study listed under a particular term. This is inevitable since the total number (106) of studies indexed is still relatively small. As this index develops into a full fledged database with the inclusion of more records, the terminology will continue to accurately reflect the variety of information available for use in research.

Thesaurus evaluation

As a test site, we were asked to comment on the thesaurus structure, the ease with which it could be used, clarity of terminology and presentation, and applicability to the subject matter being indexed. We were also asked to include terms we would either add or remove.

Our overall conclusion was that this thesaurus is not particularly suited to indexing social science data. It has long been the case that the terminology used in the social sciences is vague and that meanings change over time. That this particular thesaurus did not adequately deal with this problem is not an indication of its overall utility. However, the thesaurus did not allow us to properly describe the content of data files. A significant number of terms were completely inappropriate to the type of information contained in data files. (Sex tourism, chivalry, lookism, pronoun envy, etc.) It may be that a social science thesaurus for indexing machine-readable data needs to be developed.

At the same time, we will maintain and add to the index we've created. We will continue to use the thesaurus as a basic source, while augmenting it with additional and more precise terminology.

We decided to choose single terms for our index and did not use the delimiters. This is because the delimiters did not include terms that would be useful for the subject focus. For instance, it would have been helpful to include delimiters such as attitudes, behavior, perceptions, patterns, participation, measurements, and scales, among others. We also decided to use single terms. We had some philosophical reservations as well on terms considered sexist but heavily used in social science research terminology.

Production of the index

After indexing, the main focus of the project was to produce a printed product to be circulated among interested researchers and campus departments. Future plans include the development of an automated search and retrieval mechanism. We felt that more people would be better informed about the index if printed materials were available initially.

The final product is an index in three parts. First, there is a section of abstracts. The entries are alphabetically arranged by study title and sub-arranged by principal investigator. Each record contains the record number, study title, principal investigator(s), and an abstract. The abstracts were written consistently to contain the same elements for all entries, including a short description of the study, major topics covered, population and geography studied, inclusive dates, the number of cases or respondents, and the number of variables. A full bibliographic citation was not included because we maintain that information for our entire collection in a separate on-line catalog.

The second section is the subject index. Terms are arranged alphabetically, and study titles and record numbers are listed alphabetically under each term. Both study title and record number may be used to find more detailed information in the abstract section of the index. The subject terms section will be the most heavily used portion of the index and can be easily browsed by users.

The last section of the index contains an alphabetical list of principal investigators. Each entry is followed by the record number(s) with which the principal investigator is associated. Often a data collection has more than one principal investigator. Each of the principal investigators is included separately in this index.

In addition to the three sections, the index contains an introductory section explaining content and instructions on the use of the index. Some of the studies we indexed are part of an ongoing series of data collections, or are distributed by major archives. Since our holdings include only a selection of data available, we included a description of these archives or entire data collections in the introduction. This will alert users as to the availability of additional sources of information.

dBASE III technical specifications

The dBASE III programs are based on programs developed at Cornell University to produce the New York State Data Files Abstracts. They were adapted for the particular circumstances of the index, and then generalized to accommodate the different requirements that might arise in the production of other indices.

In adapting and generalizing the programs, the philosophy was to make the system user friendly and easy to use, and yet flexible enough so that major programming changes would not be necessary to produce indices on other subjects. Given the choice of having the computer or the user do something, whenever possible the computer would do it. For example, the program prompts for the number of characters per inch to be printed to support different types of printers, and each section may be printed individually, with a prompt for the starting page number. As a result, the programs are neither particularly efficient nor fast, and although a hard disk isn't absolutely necessary, it is highly recommended.

The record structure contains two types of fields: Those that are printed, and those that are used for internal processing by the programs. Fields that are printed include title, principal investigator, abstract and index terms. At the time a data base is created these fields can be given any name. The program will prompt for the names associated with a particular data base. The fields used for internal processing are GTERM, which must be alphabetic and set to the length of the index terms fields, and ACTNUM and STRUNUM, which must be numeric and each four digits wide.

Other than the length of the title field, as defined at the time the data base is created, there are no limitations on what can be entered into this field.

The principal investigator field may contain more than one name, in which case the names must be separated by a semi-colon (;).

To overcome the 254 character-per-field limitation in dBASE III, and to accommodate longer abstracts, multiple abstract fields may be defined. They must have the same prefix followed by a sequential number starting with 1. For example, if the prefix is 'AB', and there are two abstract fields, the fields should be named 'AB1' and 'AB2'. The data may be entered as if the multiple fields were a single field. The one restriction is that if abstract fields after the first field start with a complete word, that word must be preceded by a blank. If there is only one abstract field, it still must have a '1' following the prefix.

The naming convention for the index term field is the same: a common prefix followed by a sequential number starting with 1. Again, even if there is only one index term field, the prefix must be followed by a '1'. The program will prompt for the numbers of both abstract and index term field(s).

The records need not be entered in any particular order. Each time the program is used the records are sorted into the abstract order. The abstract portion of the index is arranged by title and sub-arranged by principal investigator. Leading English articles are bypassed in the sorting. In addition to the fields present in the record, a study number is included to be used as a reference for the subject and principal investigator portions of the index. The study number is a sequential number starting with one that is assigned to the sorted records. An option set at the time the program is run is whether to also print the record number of the unsorted file. The unsorted file number would be included to facilitate the editing process. Before printing an abstract, the program will determine if it will fit on the page being printed. Consequently, abstracts are not broken between pages, and a maximum number of abstracts are displayed on each page.

Included as an entry under each index term are the sorted study number and the title of the data file. We decided that when the subject index was being accessed it was likely the user would not know what he/she was looking for, and that we should provide the title to minimize going back and forth between the abstracts and subject index. One problem that came up was that we underestimated the length of the longest subject heading. Although we could have redefined the field length, to save disk space these long index terms were abbreviated and program converts them to the full form at printing. Abbreviations and corresponding spelled-out forms would have to be added to the program as they occur. The convention used to indicate an abbreviation was to make the last character of the index term field a period.

The principal investigator index entry consists solely of names followed by the sorted study number. We judged that users, when accessing the principal investigator index, user would know what they were looking for and the title of the data file would not be necessary.

Acknowledgements

This project was supported with funding from the Librarian's Association of the University of California, Research and Professional Development Fund.

We would like to thank the staff at the Cornell Institute of Social and Economic Research, for sharing their dBASE III programs, and giving us our start.□

Bibliography

dBASE III: user manual. Culver City, CA: AshtonTate, 1984.

Gammell, William J. "The Roper Center: development of a new organizational model for data access". Paper presented at the IASSIST Conference, Itasca, Illinois, February 8-11, 1978.

A guide to the data resources of the Henry A. Murray Center of Radcliffe College. Cambridge: Radcliffe College, April 1984.

Guide to resources and services, 1986-1987. Ann Arbor: University of Michigan. Inter-University Consortium for Political and Social Research, 1986.

Handley, Cheryl. "The Roper question retrieval system: an update". Paper presented to the

IFDO–IASSIST International Conference, Grenoble, France, September 14–18, 1981.

Index of archival holdings. Berkeley, CA: University of California, Berkeley. State Data Program, [1985].

Liskin, Miriam. Advanced dBASE III: programming and techniques. Berkeley, CA: Osborne McGraw–Hill, 1986.

Mochmann, Ekkehard; Rolf Uher, and Roy Omond "ZAR An Integrated Retrieval and Analysis System". Paper presented at the IFDO–IASSIST Conference, Grenoble, France, September 1981.

Database documentation applied to the RAND Medical Outcomes Study

by Lisa Stewart [1]
William H. Rogers

Introduction

This paper describes a schema for documenting variables in a large-scale social science survey. In particular, we describe this schema as it is applied to the Rand Corporation's National Study of Medical Care Outcomes. It serves as a basic set of working instructions for documentation. The purpose of this effort is to create a database that will enable primary and secondary researchers to locate data and comprehend the scope of the database.

The system of documentation discussed here is based on a system developed by the Language of Data Project (LOD), which is supported by a grant from the System Development Foundation. Rand initially offered to be a test site for the portion of LOD that relates to documentation. However, Rand's needs turned out to differ substantially from the use for which the available materials were designed. As a result, the LOD materials were adapted to meet Rand's immediate needs.

The material with which we began represented the early LOD thinking in Dolby and Clark (1982), Dolby (1983), and developing versions of Clark (1985). Consequently the adaptation implemented at

[1] Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Vancouver, British Columbia, Canada on May 19–22, 1987. Authors may be reached at the Rand Corporation, Santa Monica, CA 90406. The Medical Outcomes Study is sponsored by The Henry K. Kaiser Family Foundation, the Robert Wood Johnson Foundation, and the Pew Foundation.

Rand does not include much LOD structure developed since; this may be incorporated later. Meanwhile, Rand has developed extensions in other areas relating to the tracking of information. What is described here is the system as it has been implemented at Rand. For a more complete description of LOD structure see Dolby, Clark, and Rogers (1986) and other LOD materials.

The National Study of Medical Care Outcomes, also known as Medical Outcomes Study (MOS), is a multi-year panel study of health care process and health outcomes in various systems of medical care (e.g. fee for service or health maintenance organizations) and various specialty groups. It is intended to be a data resource for health policy research for several years.

Part 1 - Structural elements

I. Overview

General documentation needs

The MOS consists of interrrelated written questionnaires given to patients and their clinicians, telephone interviews, personal interviews, and laboratory reports.

A complex study of this kind produces two kinds of documentation need. The primary user (the analyst), who is knowledgeable about the study, needs a way in which to rapidy locate those data which are important to him or her, to distinguish these data from other similar data, and to be aware of peculiarities discovered by colleagues. The secondary user, who is unfamiliar with large parts of the study, needs to be able to place particular data in the context of the whole effort.

The two roles merge when data are exchanged among researchers. Thus, when the questionnaire analysts study the lab data, they need to place the data in context.

Variables

A database is structured as a table with rows that represent the subjects of the study and columns that represent facts about the subjects. The columns are called *variables.* A *variable* is a collection of observations on the same phenomenon for a set of cases. For examples, "income" is a piece of information collected on all patients in the MOS.

Variables can be divided into two types: 1) a *raw variable* is the unprocessed datum that is collected in the field. In MOS, this is obtained either through self-administered questionnaires or through telephone interviews conducted using a system called CATI — Computer assisted telephone interviews, or 2) a *derived variable* which is computed by logical or arithmetic operations performed on one or more variables, usually calculated on a computer.

Documentation of raw variables versus derived variables is fundamentally different. Raw variables are backed up by traditional documentation such as the questionnaire itself, coding specifications and

CATI scripts. In addition, questionnaire designers created "concept keys" that describe groups of closely related variables. Although these documents are voluminous, they serve primary user needs well. Secondary users involved with details need to invest a great deal of effort to understand these materials, but once they do, the information they require is there.

For derived variables, no traditional documentation exists. There is no clear statement of how the variable was computed other than a computer program which is likely to be physically and intellectually inaccessible to the secondary user. This is particularly problematic because derived variables represent the best summary of the analysts who were in the best position to understand the overall context. Thus the secondary user is often faced with the choice of using a derived variable he doesn't fully understand or deriving one himself which frequently requires a large investment if it is to be done successfully.

In the Medical Outcomes Study we focus on the documentation of derived variables and use existing sources of documentation for raw variables. This compromise is a budgetary one. While it would be nice to have a uniform summary of raw data that would help the secondary user, it is expensive to produce because there are plenty of raw variables. On the other hand, derived data are more heavily used and have no viable alternative sources of documentation.

Nevertheless, the Language of Data is based on a documentation scheme that works for both raw and derived data. This document will also describe how raw variable documentation would be done.

Data and Description

The idea behind a database documentation system is to address both levels of need, which may coexist in a particular user of the data. As a vector of numbers, a variable does not mean anything in itself, but acquires meaning from the environment that surrounds it. The first task is to document that environment, and the second task is to supply essential information for finding and using the variable.

First, we want to describe the *context* of the variable. How does it fit conceptually into the overall study? What is the sample population to which the variable refers?

By *content* of the variable we mean the conditions or qualifications that describe the datum beyond just its numerical value. To what does the variable refer? When does the measurement apply? Where? How? According to whom? As a simple example, while the numbers representing "income" may be readily accessed, the information regarding how, when, why and where the information was collected may or may not be readily available. This documentation schema *systematically* records this contextual information for each variable. These facts are usually known to the survey designers but may be unclear to secondary users.

Finally, we *identify* the location of the variable in the computer system and give other operational information needed for day-to-day use of the data. Identification elements are directed to analysts and other primary users.

In addition, once the data have been collected and the analyst works with them, he/she acquires an understanding of biases or unusual characteristics. All too often in the past, this information has been carried around in a person's head or buried in a mountain of paper. Similarly, when creating derived variables, information regarding the treatment of, or experiences with the data has rarely been captured so that others may understand one's intentions, orientation, or explanations of the variables. In this documentation system, we attempt to correct this by *systematically* recording information so that the context of the variable is computerized along with its numerical value.

General documentation approach

The steps in documentation can be thought of as two-fold. First, there is the actual documentation of the data, which is carried out on a full-screen word processor. Secondly, this information is reformatted into different types of reports by means of a rather sophisticated reporting program. This paper focuses on the first half of the process, serving principally as a working manual for documentation.

The first step consists of identifying and recording entries for the essential elements needed to describe the variable. An *element* or *descriptor* is a basic piece of information that answers what, when, where, why or how issues for each variable. At Rand we refer to the categories of descriptive elements as a *template* or *descriptor set* and sometimes refer to the process of documentation as "templating". The importance of this term is that we systematically record the same information for every variable in the study.

After all the data have been documented, the descriptive information is transmitted to a computer-based reporting system for locating variables and displaying their context. This system is analogous to a library system for finding books. Over the long term, we hope that LOD will have the resources to develop a computer reporting system that can compute symbolically and linguistically the contextual information as well as one now does numerically the numbers.

II. Introduction to elements

The categories of descriptive information which we use to document the variable are grouped into *context*, *content*, and *identification* elements.

The *context* elements are *subject* and *topic*. *Topic* in turn is organized into *main topic* and *subtopic* in a hierarchical (tree) structure. In a study such as the MOS, the *subject* describes the set of cases to which the variable applies. The *topic* describes where the variable fits into the MOS Concept Outline.

Next there are five major content elements: *observer*, *matter*, *function*, *space*, and *time*, and two associated content elements: *aspect* and *domain*.

Finally, the identification elements (which include categories developed at Rand) provide historical and access information about the variable. They are *variable name*, *location*, *source*, *prior source*, *status*, *creation data*, *analyst*, *and footnotes.*

Context Elements

(Indicate how each variable fits into the overall study)[2]

SUBJECT: The set of cases to which the contents of the variable refer. (All patients, all clinicians, patients with diabetes, etc.)

TOPIC: The categories under which the contents of the study are organized[3]

MAIN TOPIC: In the MOS, the top-level study concept under which subtopics are organized (Tracer Conditions, Style of Care, Provider Characteristics, etc.)

SUBTOPICS: In MOS, the study concepts under which groups of related variables are organized (see the MOS "Concept Keys")

Content Elements

(Descriptive structure of each variable)

OBSERVER: The person supplying the values of the input variables; in a survey, the respondent category (doctors, patients)

MATTER: The objects involved in the event discussed

FUNCTION: The nature of the event, the activity or state being discussed

SPACE: Where the event occurred or the location to which it applies (physicians' offices, at home)

TIME: When the event occurred (1986, 1 month prior to screener fielding, patient's lifetime); not synonymous with data-collection date

ASPECT: The specific characteristic being observed

DOMAIN: The nature of the values, a description of the units and/or range of measurement (1=, 2=; 1,...,100)

[2] SOURCE and OBSERVATION DATA have an important bearing on the context and may become context elements for the secondary user (to be included here).

[3] LOD structure includes two extensions not covered here. One is classified description that starts at the topic level, as a guide to the levels below. The other is an extension of the topic structure to the local topic for each variable, represented by one of the content elements in the descriptive structure. For a discussion of these extensions see Dolby et al. (1986) and Clark (1985).

Identification Elements

(Access and historical information)

VARIABLE ID:	An 8-character alphanumeric code name for the variable
DATA LOCATION:	The data set in which the variable is stored
ANALYST:	The analyst, the author of the variable
PROG SOURCE:	A pointer to the program defining the variable
PRIOR SOURCE:	A pointer to the input variables (e.g. SD1MI01-SD1MI05) and/or documents used to define a derived variable (MOS memos, journals)
VAR CREATN DATE:	Survey date or date the program was run to create a derived variable
VAR STATUS:	Administrative indicator of the variable's stage of development: PROPOSED (no data yet), PRELIMINARY, SEMI-PERMANENT, PERMANENT

FOOTNOTES:

1. Variable construction (scoring, input variables, missing data rule)
2. Analyst's comments (reason for creation, experiences with variables)
3. Documentor's comments (notes on series of/relations among variables)
4. Administrative variable noting age of template information

III Templates: elaborated definitions and working rules for raw and derived variables

Context Elements

1. SUBJECT: The set of cases to which the contents of the variable refer. (All patients, all clinicians, patients with diabetes, etc.)

   In the MOS concept outline, this is actually the highest classification level for the variables.

2. MAIN TOPIC: is a top-level survey concept under which subtopics are organized.

   A variable should be classified according to the topic and subtopic which best describe that variable. These should be chosen from an approved list.

3. SUBTOPIC: is a major survey concept; related items (variables) grouped together measure a particular concept.

   The main topic / subtopic structure defines the variable's *raison d'etre* from the study's viewpoint. For example "depression" is a subtopic of "mental health" which in turn is a subtopic of "general health". One should be careful in choosing an entry; occasionally a variable will measure something other than what appears on the surface (e.g. subtopic: Socially Desirable Response Set).

   In the MOS survey, these concepts are laid out in the Project Overview in the Figure called, "Conceptual Framework and Key Study Variables."

Once one has a working knowledge of how the various concepts fit into the study's overall organization, one may add the content elements.

Content Elements

1. OBSERVER: person through whose eyes the real-world phenomenon is viewed; in the case of a derived variable, the person who specified the value of the input variables

   In MOS, this will usually be the patient but may be the patient's physician, an MOS laboratory (for a blood test), or other.

2. MATTER: the object(s) observed; whatever objects the data apply to.

   This identifies who or what the question, derived variable, or concept is about. Matter MUST be a physical object (as opposed to a state-of-being).

   In MOS, this will be the patient, a clinician, or a patient-clinican dyad. In some cases you may encounter possessives, such a "(patient's) primary physician". For MOS variables, the stub (the objects in the sample) will always be part of matter.

3. FUNCTION: is the event happening to the object or the state-of-being of the object.

(Read this carefully) Where matter is the object to which the event is happening, function is the event itself. Where matter is the object to which the state-of-being refers, function is that state-of-being itself.

CAUTION: The word "function" is a technical word in classified description and (in a different meaning) a topic of the MOS (e.g. physical functioning, etc.). Don't confuse topics of MOS with the content categories.

4. SPACE: where the event occurred (e.g. physicians's offices, at home) or where the topic applies (e.g. continental U.S., universal).

   If there is no specific place, the answer is often "universal." Also, don't make inferences regarding space because it can be something different from what one would assume. For example, if a patient had a baby, it would be a logical assumption that the event had occurred in a hospital. This might hold true for 98% of cases, but it could also be that 1) baby was delivered at home by a midwife, or 2) baby was delivered at a clinic, or even 3) mama didn't make it to the hospital on time and baby was delivered in a taxi. The only way one could assume the birth took place in a hospital would be if one were in the hospital sampling women who had just given birth there, or if the hospital were specifically mentioned in the question.

5. TIME: when the events described occurred (e.g., 1986, the first year of participation, any time before Sept. 1, 1985); the time of the datum.

   Most often the value of TIME is the date of observation (i.e., the data are current as of the time the questionnaire is filled out). In MOS, the same instrument is sometimes fielded at multiple times. The various fieldings become discrere variables in the database and must be distinguished by their variable IDs. For raw variables, the first three characters of the name are reserved for the source instrument and fielding time (e.g. AA1DRG01, AB1DRG01).

   Sometimes TIME will differ from the date of observation. For instance, in 1986, a questionnaire might be eliciting information about 1985 income. In this case, observation date would be 1986 and TIME would be 1985.

   EXAMPLE: (for occupation or education) Time of observation
   EXAMPLE: (for 1974 income) 1974

   Sometimes the time of observation is another variable in the dataset. If so, the variable name should be given.

   Note: The archivist should check at the beginning of a series of questions for instructions regarding time. Time is frequently indicated there rather than being repeated for each question in the series.

6. ASPECT—the specific characteristic being observed.

   This should be an elaboration of matter, or function, or one of the other descriptors. Examples: frequency of, annual cost, patient's evaluation of ...)

The entry should describe the elaboration as well as that which is being elaborated (the antecedent). The antecedent specifies one of the other elements in parentheses followed by the entry for that element.

Aspect has the form "xxx of (y) zzz" where:

— "xxx" is frequently amount, degree, type or name corresponding to whether the measurement is continuous, ordinal, categorical, or an identity.

— "y" (in capital letters) is most frequently F for function or M for matter, and occasionally T for time or S for space.

— "zzz", the referent, is the entry for matter or function.

For example: "Amount of (F) Mental Health Distress", or "Subspecialty of (M) the physician."

```
"Amount of (F) Mental Health Distress"
   ^        ^    ^
     Aspect       Function
```

— DOMAIN: is a description of units and/or range of possible values.

For example: mm of mercury, Mental Health Index (MHI) scale points 0(poor)–100(good), A=excellent B=good, -2=don't know.

The data domain presumed and reported by the analyst should be distinguished from the actual domain observed in the data. Unless otherwise stated, we are referring to the domain presumed by the question.

For example, a questionnaire may have response choices that are printed on the questionnaire but are never used by any respondents in a given administration of the questionnaire. We might never observe an American Indian in the study even though we have a response category for American Indian. The existence of the possibility of such a response affects the interpretation of the rest of the choices (e.g., we are safe in assuming that Caucasian does not mean American Indian).

Likewise, a respondent may write his/her own answer in the margin of the questionnaire, and the survey coder may be tempted to treat this as a new code. He/she should not do this. Since the write-in response is not one of the given choices, it will not be chosen by a typical respondent. Consequently, the write-in response does not reflect on the choices of the typical respondent. For example, a person might write in the response "transvestite" in response to a question about gender. Some other transvestite might respond "male." So we can't make inferences about the spontaneous category since it was not considered by other respondents. The given categories are the ones that we want to record.

The stated domain will be used for automatic compatibility checks—a standard method of detecting errors in the data.

SYNTAX: The entry should be machine parasable and not a graphic picture of the answer layout. It is presumed that the description adequately represents the choices available to the respondent. Lengthy data codes (e.g., FIPS county codes) will not be listed in the template format library. Instead there will be an example of the format, then a reference to hardcopy translation.

Unlike any of the other facets, the type of value is recognizable by the syntax used. Conventions are as follows:

Syntax used:	*Examples:*
Discrete Variables	
-- Categorical responses specific sets of integers	*1=poor, 2=fair, 3=good, 4,excellent*
-- Boolean logic	*0=no/false, 1=yes/truee*
Continuous Variables	
-- Rational (grainy, measured to limited number of decimal points) NBS format	*0 (.01) 999.99 meters*
-- Integer range (special case of above)	*1(1)5 OR 1,2,3,....500 units*
-- Real (without grain information (varying number of decimals after decimal point; usually computed)	*[1;infinity] years -or- [-inf;+inf] std units*
-- Date types, usually in the SAS form MDY date (MDY) (see SAS manual for all forms)	

Identification Elements

1. VARIABLE ID: is an 8-character alphanumeric name reserved for this variable.

   In MOS, variable IDs are limited to 8 characters in order to fit SAS programming conventions. Two or three characters are reserved to indicate the dataset to which the variable belongs.

   For raw variables, names begin with a 2-character instrument code (Moser's, see /a/p/reference/instr.codes) and a 1-field version/administration number. The remaining 5-characters are assigned by analysts. This portion of the name comes from the "measures keys" tables analysts maintain and is supposed to be unique (e.g., PP1HLT01 is the same question as SP1HLT01) within the study.

   For derived variables, the first 6-digits are the name, followed by (where appropriate) a 2-digit code indicating the source of the input variables. The 2-digit code is stored in /a/p/reference/derived.suffix.

2. DATA LOCATION: is a pointer to the dataset in which the variable is stored.

   In MOS, raw variables will be stored, by instrument, as a SAS dataset on WYLBUR. The same applies to derived variable datasets, except that they will be combined into either a patient or clinician database (not instrument specific). Details are worked out in accordance with programmer needs.

3. SOURCE: is the origin of the data—theoretically the source document(s) defining the variable.

   This may become part of the content elements at a later date, when the source is fixed as "Rand" by virtue of publication of study results.

   In the meantime it is very important to track information internally within the project. There are two senses in which this must happen:

   PROG SOURCE: is the computer program in which the variable is defined (needed to gain access to the unambiguous definition).

   PRIOR SOURCE: is the documents used to define a variable (MOS memos, papers, and previous studies). For raw variables, it is most typically an MOS memo describing the variable. For derived variables, the prior source is an original source (or sources) where the data used to construct the derived variable were defined. This will usually refer to a questionnaire or (in complex cases) several questionnaires (from the present study) and will have, tagged onto the end, citations to any previous study.

Table 1

Raw Variables	**Derived Variables**
Source—Typically, a MOS memo. The source could be a previous study if that study stands alone to define the variable.	Source—Location of programming statements used to define the Variable.
Prior Source—If a MOS memo draws from a previous study for variable definitions, reference that study, or other outside source, here.	Prior Source—The Source (and if there is one the Prior Source) from the raw variable.

SYNTAX: See APPENDIX D for specifics.

4. ANALYST: For derived data only, the name of the analyst creating the variable. The analyst is considered an author and it is the anlyst's view that produces the description. We therefore record the context appropriate to this description.

5. OBSERVATION DATE/VAR CREATN DATE: For raw data, the survey date; for derived data, the date the program was run to create a derived variable.

   The date on which a derived variable is created we call the "variable creation date" to distinguish it from the date of collection of the raw variable, the "observation date".

   Even in the case of a simple scale, the propriety of a given decision to include certain items and exclude others is apt to become dated.

   In filling out entries for observation date (and time), one pitfall the archivist needs to avoid is omitting instructions regarding a series of questions. For example, if there is a series of questions for which the patient needs to think back over the past month, (e.g., "For the next series of questions, please consider your feelings during the past month."), be sure to take that instruction into account for all the variables it applies to.

6. VARIABLE STATUS: Administrative indicator of the variable's stage of development: PROPOSED (no data yet), PRELIMINARY, SEMI-PERMANENT, and PERMANENT. The purpose of this is to encourage sharing of well-documented information early in the project.

   Only derived variables need this indicator. (Raw variables have been finalized by the time researchers see them. They had to be in order to be fielded. Thus, upon arrival of raw data, their status is known.)

Proposed item — analyst has conceived the variable. (The variable has a code or programming statements, but it hasn't been created yet.)

Preliminary item — analyst has created the variable, but the meaning can still change.

Semi-permanent item — analyst has verified that this variable measures the concept precisely. In practice, a variable is semi-permanent if the analyst does not expect to change the formula.

Permanent item — has not been revised for at least six months.

7. FOOTNOTES: contain information that does not fit neatly into one of the other prescribed categories or is too voluminous (such as the question text), or information the analyst or programmer wishes to supply explaining the origin of the variable. In documenting MOS variables, we emphasize that the latter, the origin of the variable, is important information that all too often in the past has remained in the heads of the analysts, thus depriving other users of a full understanding of the variable.

   The structure of footnotes is free-form and may be enhanced to meet whatever needs the analyst finds appropriate. However, in our experience the following entries have been considered valuable.

   a. For raw variables, the verbatim text of the survey question. For derived variables, a description of the variable construction (by means of: scoring description, input variable listing, missing data rule).

   b. Analysts comments: For raw/derived variables, a record of analysts' experiences with the variable. These should be collected by the system and appended to the documentation. For derived variables, notes regarding how/why the variable is constructed. Frequently includes the reason for the variable's creation.

   c. Documentor's comments frequently tie together related variables (e.g., This is a series of 3 utilization variables increasing in complexity: PUTIL3 is derived from PUTIL2 which in turn is derived from PUTIL1).

   d. Documentor's administrative variable indicating age of the information contained in the template and whether analyst has reviewed the template.

PART 2 - DOCUMENTATION MANUAL

I. Parallelism Among The Study's Topics/Concepts or Get To Know Your Data

The most difficult aspect of this documentation system is choosing the right level of specificity for entries in the *content* descriptors. If the level is too general, there will be no visible distinctions among large numbers of variables. If it's too specific, both documentor and readers are swamped with detail. One can arrive at the right level only by considering a set of variables.

In order to maintain a structure of parallelism with any degree of accuracy, one must identify a hierarchical summary of the study's major concepts. For MOS, this was provided in a summary table in the Project's Overview Document (Ware, 1985). The framework of a topic classification is obtained from this document.

An archivist/documentor should conduct a thorough review of the data collection instruments to become familiar with the items in each. (This is especially important for MOS instruments which usually reflect several different measures.) Next, it is imperative to identify the major study concepts each item in the instruments reflects. Connecting specific items to concepts is essential in order to classify variables in a reasonably consistent fashion.

Overall Strategy

As stated above, probably the most difficult part of data classification is to choose the correct level of specificity. This corresponds to the right "discourse level" of a conversation. It is important that parallelism (of the level of specificity) be maintained when documenting.

If a question pertains to an event, it is relatively easy to choose entries: matter refers to the object(s) involved in the event and function specifies the event. If the question pertains to a state-of-being, the matter is usually clear but the function is not.

For example, a question such as "During the past week how often did you feel blue?" could be construed as a measure of health, a measure of mental health (a subtopic of health), a question about depression (a subtopic of mental health), or a question about "blueness." Depression would be the best function description, and "Amount of (F) Depression" would be the corresponding aspect. Without the knowledge that this question really refers to depression, you might describe the function as "blueness." The other choices (mental health and health) are poorer because they lead to complicated ASPECT descriptor entries.

Varying Number Of Levels In The Hierarchy

Even with knowledge of the topic structure, there will be times when it is difficult to determine the right level of specificity. One aid that we at Rand have found helpful is to make an outline of the instrument being documented. This lays out graphically the structure below the subject/main topic/subtropic level.

We have found that sometimes it is not possible to use the best description of the variable (as described in the aspect) and still have the hierarchy "fit" logically. There is no getting around the fact that the number of levels between topics and items varies.

Consider, for example, a variable describing physician's date of graduation from medical school. It is part of a series that includes sociodemographic (e.g. age) and other education items. The items appear to be parallel in the questionnaire, but to make them parallel in the documentation, we would have to work education into the descriptors along with the additional specifier (date of graduation). If we didn't, there would be no logical connection between demographics and year of graduation (from what?). Our solution: add education as a sub-subtopic.

VARIABLE ID:	CGRADYR
DATA LOCATION:	R.R5500.A4195.CAMADV
PROG SOURCE:	/a/p/programs/ama.pgm
PRIOR SOURCE:	AMA Physician Masterfile (1985), Rand Data Facility DB 303, distributed by American Medical Association, obtained for Rand: 1986
STATUS:	Preliminary
Var CREATN DATE:	04/11/86
ANALYST:	Bill Rogers
OBSERVER:	Clinician
SUBJECT:	All Clinician's enrolled in the MOS Panel Study
TOPIC:	Clinician Characteristics
SUBTOPIC:	Demographics/Socioeconomic status
SUB-SUBTOPIC:	Education
MATTER:	Clinician
FUNCTION:	Graduation
ASPECT:	Year of (F) Graduation
SPACE:	Universal
TIME:	Universal
DOMAIN:	Date (MDY)
FOOTNOTES:	(1) SCORING: Raw item recode INPUT VARIABLES: BS1GRDYR MISSING DATA RULE: No imputation applied (4) 10/15/86 Template reviewed by Bill Rogers

II. The Technicalities Of Making Templates

We have found that the best technique for documenting a large dataset is to fill in as many common entries (e.g. data location) as possible, then copy the template using a full-screen editor from variable to variable and fill in the entries as required. Most raw data entries change little from variable to variable. For both raw and derived data this saves repetitive typing.

Strictly speaking, the term "templating" describes this activity.

One should distinguish between input format and display format. The input format is designed for ease of entry and its ability to be manipulated by string processing programs such as "awk". This is the input format:

Unreported Template

In its unreported form, the template looks like:

```
VARIABLE ID:
DATA LOCATION:
SOURCE:
PRIOR SOURCE:
STATUS:                 (derived only)
OBS DATE:               (or VAR CREATN DATE for derived variables)
SUBJECT:
TOPIC:
SUBTOPIC:
MATTER:
FUNCTION:
SPACE:
TIME:
ASPECT:
DOMAIN:
FOOTNOTES:
```

Notice that there is a difference between raw and derived variable templates. The derived variables template has three more elements than the raw template—STATUS, ANALYST and PROG(ram) SOURCE.

Syntax Conventions

Certain syntax conventions must be maintained so that the data may be parsed (separated and read) by a computer program which will report the data in various forms.

The syntax conventions are:

1. Type in entries after the colon.

2. The text may wrap around to the next line, but should not start in column 1.

3. Leave a blank line between variables.

4. Separate fields by commas (there may be more than one entry per facet).

5. Use special syntax for a particular facet where specified, e.g., for domain

6. Type dates as: mm/dd/yy

7. Orthography requirements:

a. Use the same capitalization and punctuation for entries as one would in English (e.g. capitalize important words).

b. For names, use the standard conventions, e.g. /a/p/programs/ama.pgm, R.R550.A4195,CAMADV, etc.

c. Capitalize all letters in names of variables.

8. If there is no entry for a facet, type "None", so we'll know it has been considered.

9. When there is more than one entry for a facet, separate them by semi-colons.

. Dealing With Uncertainties

If you don't know an entry at all, DON'T GUESS. Leave it blank if you are completely baffled and ask for help. If you have a preference but are unsure, start with "(?)". Classifications may be revised later by the analyst who invented the questions or variable (and who presumably has a better perspective on the intent).

Special Instructions For Derived Variables

When making derived variable templates note the name of the input raw variables (in parentheses) wherever possible, For example, note the variable names in parenthesis in the following template:

VARIABLE ID:	PELIGGRP
DATA LOCATION:	R.R5500.A4195.PDERIVED.SAS
PROG SOURCE:	/a/p/programs/derived5
PRIOR SOURCE:	MOS memo123
VAR CREATN DATE:	Screener date (PSCRND)
ANALYST:	Bill Rogers
OBSERVER:	Patient, Clinician
SUBJECT:	All Screened Patients
TOPIC:	Survey Administration
SUBTOPIC:	MOS Panel Study
MATTER:	Patient
FUNCTION:	Eligibility
SPACE:	Universal
TIME:	Screener date (PSCRND)
ASPECT:	Eligibility of (M) patient
DOMAIN:	0 = No hypertension; 1 = Hypertension
FOOTNOTES:	(1) SCORING: Look at both patient's, and clinician's reports of hypertension; INPUT VARIABLES: SP1HYPO1, SD1HYPO2 MISSING DATA RULE: Recoded to missing if either input variable value is missing.

(4) 10/15/86 Template reviewed by Bill Rogers

Also, please notice that derived variables frequently have more than one entry for certain facets. For example, a derived variable may combine the viewpoint of both the clinician and the patient, so both would be entries for the observer facet.

III. Practice and Initiative--Raw Variables

Now we will make a first attempt at descriptive classification at the raw variable level. Specifically, how are the elements connected to reality?

Step-By-Step Raw Variable Classification Example

Helpful Hints

Let's begin by looking at a set of requirements for templating, which need to be kept under consideration in order for the pieces of the puzzle to fit. For example, consider the following issues, in order:

- - the aspect must originate from (or reflect) the domain
- - the aspect must refer back to either matter, function, (or in a few rare cases, to time)
- - function must directly connect with subtopic (or sub-subtopic)
- - within MOS, subtopic and topic fit in with the overview concepts list (Overview, Conceptual Framework and Key study variables) with topic above it.

Keeping these requirements in mind, in the following order, let's determine the entries for MATTER, DOMAIN, ASPECT, FUNCTION, TOPIC, SUBTOPIC, SUB-SUBTOPIC (where necessary), and then the remaining elements.

Specific Example

For a detailed example, let's consider a question that asks: "How many different drugs are you taking for your high blood pressure?" Document the question by doing the following:

1st) Record matter:

MATTER: Patient

Choose patient because we know the question is asked in regards to the patient.

2nd) Record domain:

DOMAIN: 1=none, 2=one, 3=two or more

This is taken directly from the questionnaire. We do this right away since we know that the aspect has to reflect the nature or quality of the response set.

3rd) Record the first half of aspect:

ASPECT: Number of ....something

Choose this since all of the above in domain indicate a specific quantity, a number.

4th) Record the second half of aspect — the antecedent (the thing the aspect refers to):

ASPECT: Number of (F) Drugs taken for Hypertension

"Drugs taken" must be the answer because that is what "Number of" refers to. We complete the phrase by adding "for Hypertension" since we know that is true and relevant here.

Notice that aspect can be very wordy. It is the element that pulls the other elements together.

Sometimes the antecedent refers to another element, commonly the matter. We know this is not the case here because "Number of (M) patient" is not only incorrect, but also illogical.

5th) Record function:

FUNCTION: Drug Usage

Since the second part of the Aspect must refer back to another element (usually matter or function) and the value of matter is already "patient", function is almost predestined to be Drug Usage (a more formal way of saying "taking drugs").

Now one might consider filling in some of the higher-level entries which indicate how this variable fits into the overall scheme of the study, making any adjustments needed along the way to make the lower- and higher-level entries meet in a reasonable, logical fashion.

6th) Record main topic (highest major study concept):

MAIN TOPIC: Patient Characteristics

This variable could have been created to measure Utilization under the topic of "Process of Care", but the analyst creating the variable is the MOS M.D. who is responsible for describing the patient, so "Patient Characteristics" is the best choice.

This variable can also occur as a patient outcome measure but because it is asked at the beginning of the study, we know it is part of initial "Patient Characteristics". Toward the end of the study, the documentor should check with the analyst to determine whether or not the variable has been repeated, and if so, add "Patient Outcomes of Care" as a second entry for this descriptor element.

7th) Record subtopic (major study concept):

SUBTOPIC: Disease Severity

The only way to know for sure which subtopic is correct is to personally ask the analyst what his/her intent is in asking the question. This entry could have easily be mistaken for a measure of utilization instead of disease severity.

8th) Create sub–subtopic to make a logical connection between SUBTOPIC and FUNCTION:

SUB–SUBTOPIC: Hypertension

As it stands right now, the distinction between "Disease Severity" and "Drug Usage" is too weak to stand by itself and does not properly reflect the content of the question. Here is where we need to make adjustments by adding another level of specificity to the template. For the SUB–SUBTOPIC, choose hypertension since that is what the drugs are taken for and how the analyst is measuring the disease severity. Now the link between these three levels is very logical and clear to all analysts.

9th) SUBJECT—identify the pool of people to whom the variable applies

SUB–SUBTOPIC: 1/2 of screening patients

We choose the above since we know that 1/2 the patients screened filled out SP2 and the other half filled out SP1.

Now that the hard part is done, let's round things out by filling in the easier, more obvious entries:

10th) LOCATION — R.R5500.A4195.PDERIVED.SAS

dataset in which variable data resides; supplied by analyst/programmer

10th) VARIABLE ID — SP2PHYPO5

supplied by programmer or analyst

11th) LOCATION — /a/p/datasets/SP2

dateset where variable data resides: supplied by programmer

12th) PROG SOURCE — /a/p/programs/pdv.part1

13th) PRIOR SOURCE — Shelly Greenfield, "Determining Disease Severity", MOS memo 515, 10/10/85

MOS memo cited first, follows; see Appendix D for citation conventions

This could also be a journal article, another research study or for MOS, the raw variable documentation the questionnaire designers created, called "MOS Concept Keys".

14th) OBSERVER — Patient

We know this becasue it is the patient who is filling out the form

15th) OBS DATE — Screener Week

This is the date of data collection.

16th) SPACE — Universal

Number of drugs taken is constant regardless of where the respondent is.

17th) TIME — Screening Date

The time frame is the same as the date of observation by default since the question is asked in the present tense.

Sometimes the archivist has to look at the beginning of a series of questions for directions on time.

18th) FOOTNOTES — (1) How many different drugs are you taking for your high blood pressure?

Question text goes here (no graphics).

(2) Analyst assumes that worse cases require more drugs.

Analyst assumptions are important information that the documentor should be careful to record.

(4) 10/15/86 Template reviewed by Steve Rein

An Administrative variable indicating the age of the information in the template is goes here.

As mentioned above, there is not always a single right answer for an entry. However, you should apply a uniform viewpoint across the questions so that the dimensions are consistently used. Parallelism is very important.

The resulting template is:

VARIABLE ID:	SP2HYPO5
DATA LOCATION:	/a/p/datasets/SP2
PROG SOURCE:	/a/p/programs/pdv.part1
PRIOR SOURCE:	MOS memo 515
OBSERVER:	Patient
OBSERVATION DATE:	Screener week
SUBJECT:	1/2 of screening patients
TOPIC:	Patient Characteristic
SUBTOPIC:	Disease severity
SUB-SUBTOPIC:	Hypertension
MATTER:	Patient
FUNCTION:	Drug Usage
SPACE:	Universal
TIME:	Screener week
ASPECT:	Number of (F) drugs taken for hypertension
DOMAIN:	1 = None; 2 = One; 3 = Two or more
FOOTNOTES:	(1) How many different drugs are you taking for your high blood pressure? (3) Assumes that worse cases require more drugs to control hypertension. (4) 10/15/86 Template reviewed by Steve Rein

For a graphic example of what happens when template entries are filled in without the analyst's intent being verified, see Appendix C.

IV. Practice and Initiation--Derived Variables

Next let's learn how descriptive documentation works at the derived variable level. To understand this one needs to know how:

a) a derived variable is defined (compared with raw variables);
b) item-coded derived variables are similar to raw variables;
c) to make a derived template;
d) derived variables have special documentation concerns
e) a "status" line indicates variable stability
f) derived variables develop in the analysis process

A. Definition of Derived Variables

A derived variable is one that:

1. is not present in raw data and has been created by some kind of computation.

2. is based on an assumption made by the analyst with regards to the meaning of the variable, and is derived from the data plus the analyst's logic.

Raw data from an outside source are considered derived data for the purposes of the study being documented because the data were not generated by the home study instrument.

Derived variables vary greatly in their complexity. Simpler derivations are more numerous and require less sophisticated handling. Conversely, more complex derivations occur less frequently and require sophisticated handling.

Let's look at four cases of derivation, from most simple to most complex:

1. ITEM RECODING has two major functions:

   a. to transform the data to a more computational form (e.g. reverse scaling to align with other variables, change alphabetic characters to numeric, etc.)

   b. reduce error response (e.g., resolve multiple punches, data inconsistencies, out-of-range values, etc.)

   Commonly, simple derivation of variables is tagged onto the raw variable data processing routine.

   While these two actions are both considered simple item recoding, one is a *substantive* change and the other is a change in *form*. They need to be treated in distinct ways. See "Item Recoding and Similarities to Raw Variables" below for more information.

2. SCALES are a product of a systematic combination of individual item scores into a summary score[4].

3. COMPLEX COMPUTATIONS involve formulae that use substitution in the case of missing data (i.e., if the date of a visit is missing from a patient file, it may be picked up from the clinician file).

4. SUPER COMPLEX COMPUTATIONS involve predicting values and making reference to multiple observations in a file. They are so complex because the predicted value depends on the values of the rest of the dataset and sometimes requires use of regressions[5] (e.g., predicting age from income).

B. Item Recoding And Similarities To Raw Variables

A recoded item is harder to classify than other derived variable types because it changes variables in two different ways: 1) *substantive* changes and 2) *form* changes. Both of these are called item recoding because they are both relatively simple changes, which involve only one variable, limited programming (usually a single line) and assumptions about the variable.

[4] A scaled response is different from a scaled variable, and refers to a graded response set (e.g., excellent, good, fair, poor)

[5] A regression is a model which explains the behavior of some data given the behavior of other data.

The following situation is an example of a *substantive* change. If the analyst wants to make assumptions about response possibilities in order to limit errors, he may define age as 1) having the upper bounds of 100 (age value = 0–100), and 2) having the formula (PSCRND – PBIRTHD)/365. This would get rid of many respondent errors (e.g., all 4–digit numbers, alphabetic characters, etc.). It also introduces an assumption that no one was over 100 years old and that there were no decade errors, that people did not write in their age, etc.

On the other hand, *form* changes are actually disguised raw variables. Since the essence of the information is the same, and only the way that information is recorded is different, form–changed variables are very similar to raw variables. Mostly, this type of re–recording of information is done to standardize a measurement (e.g. make all scales conform to values of 0–100, or reverse the direction of a single scale to alighn with the other scales) for use in a scaled derived variable.

Substance–changed variables call for different classification than form–changed variables. While the analysts process these two actions at the same time, we need to document them differently. Recoded variables with no substantive changes need to be templated as raw variables and substantively changed variables need to be templated as derived variables.

C. Creating Derived Variable Templates

An archivist needs to keep the above differences in mind in order to choose whether to use a raw variable template or derived variable template for documentation. (See APPENDIX B for examples of raw and derived variables). The raw template is so similar to that of the lowest–level (item recoding) derived template, that for convenience sake, they will henceforth be referred to together. In order to make a derived template one may start with a raw variable template, and make the following changes:

a) Add PROG SOURCE, STATUS, and ANALYST to the template.
b) Change OBSERVATION DATE to VAR CREATION DATE.
c) Record variable construction information (SCORING, INPUT VARIABLES, and MISSING DATA RULE) in the footnotes.

D. Special Concerns Of Derived Data Documentation

At the data processing level, a major issue in archiving derived datasets is keeping the data in a similar stage of processing. Analysts sometimes lose interest mid–stream which: 1) ignores data which can be different from those already analysed and can cause significant variations in statistical outcomes 2) creates unevenly processed data. An archivist, in tracking the data, should look out for and note this problem for both the analysts' benefit and that of future users.

Another problem specific to the documentation of derived variables concerns capturing data in the most accurate state. For raw data, that may be either (1) when the questionnaire author writes the questionnaire, or (2) when analysis occurs and the question is re–read with a more critical eye. Interesting facts about the raw data may come to light as derived variables are created.

For derived data, there is a similar problem. The concept being derived from a given set of data may evolve over time, or the intensity of knowledge may disappear with time (people may forget assumptions made at the beginning of data collection), leading to mistaken classifications. The problem is reflected in processing: programs may be written which refer to data in one state or another, and it is difficult to tell whether the program should refer to older data or revised data.

In general, derived variables are in flux longer than raw variables and there is no equivalent to the questionnaire text to finalize the meaning of derived variables. Raw variables are fixed as of the date of fielding, whereas derived variables are always subject to change, making them more elusive to understand and document accurately.

E. Status Line To Indicate Variable Stability

For the MOS study, we decided to create an indicator to control the state of flux derived variables are so frequently in. The purpose of the "STATUS" indicator is to flag the underdeveloped derived variables so that more current meanings will be recorded later. To learn to choose which value the STATUS line will be, let's look at the analytic process and how derived variables are created.

F. Derived Variable Development In The Analytic Process

In social science surveys, *measures* are the instruments or tools used to represent concepts.

Typical stages of development measures go through are:

a) Analytic history and technical literature illustrate what research has been done in other projects.

1) *Proposal* stage - these are ideas about *what* should be measured, and thoughts about *how* those ideas should be measured; in the MOS these are spelled out in the Overview of project and in the myriad memos on measures.)

2) *Preliminary* stage - concepts are broken down into specific items and values (item wordings and groupings); both of the latter are subject to change. Tentative programming code is written and run, creating the first version of the variable.

3) *Semi-permanent* stage - items accepted, values subject to change, adjustments made to computations. This is the point at which the variable becomes static. It has been subjected to and passed measurement analysis. (For MOS variables, this would typically be standard psycho-metric criteria.)

4) *Permanent* stage - Not only is the variable static, but no one has even looked at it for six months. It has been computed for all cases in the study database.

A sample application of this process is:
1) Researcher proposes studying income; writes measures memo
2) Economist says wages is a more appropriate measure than income; writes formula for wages
3) After reviewing the first version of the variable, researcher adjusts definition of wages or makes corrections to programming
4) No further changes made to data or definition

When documenting MOS variables, we found the "PROPOSED" value the least valuable since variables at this stage are often too tentative to document. The other indicators are valuable as a mechanism for finalizing a variable's definition with the most recent information.

V. Step-By-STEP DERIVED VARIABLE CLASSIFICATION EXAMPLE

Let's keep in mind again, some criteria that need to be met in the course of templating:

— the aspect must originate from the domain
— the aspect must refer back to either matter, function, (or in a few rare cases, to time)
— function must be directly linked to subtopic
— within MOS, subtopic and topic fit in the overview concepts list (Overview, Conceptual Framework and Key study variables)

So, let us determine, first, matter, domain, aspect, function, topic, and subtopic, then the remaining elements, in that order.

For a detailed example, let us consider a question that asks: "What was your household income, before taxes, in 1985?" Document by doing the following:

1st) Record matter:

MATTER: Patient

Choose patient because we know that is who the question is about.

2nd) Record domain:

DOMAIN: [1:infinity]

This is the range specified in the questionnaire response set.

3rd) Record the first part of aspect;

ASPECT: Log of ... something

Choose this since the numbers reported in the domain are being reformatted into a log for each income level. Archivist wouldn't necessarily know this and may have to clarify when interviewing analyst about derived variable meaning.

4th) Record the second half of the aspect — the antecedent
(the thing the aspect refers to):
ASPECT: Log of (F) Income

This can be surmised by default since a "log of patient" is nonsense.

5th) Record function:

FUNCTION: Income

Choose income because we are talking about that particular characteristic of the patient and this was already determined by the antecedent of the Aspect.

6th) Record topic (top-level study concept):

TOPIC: Patient characteristics

This measure is a patient characteristic that is part of the initial description of the patient.

7th) Record subtopic (major study concept):

SUBTOPIC: Demographics/socioeconomic status

This subtopic is fairly obvious. Most researchers consider income a demographic measure.

8th) SUBJECT—identify the pool of people to whom the variable applies

SUB-SUBTOPIC: All screening patients

We choose the above since we know that this question was asked on both forms of the screener questionnaire.

Now to round things out, let's fill in the easier, more obvious entries:

9th) VARIABLE ID — PINCLOG

supplied by analyst, prefaced with a "P" to note a patient derived variable.

10th) LOCATION — R.R5500.A4195.PDERIVED.SAS

dataset in which variable data resides; supplied by analyst/programmer

11th) PROG SOURCE — /a/p/programs/derived5

location and name of program that created this derived variable

12th) PRIOR SOURCE — MOS memo cited first, a prior study's definition (e.g., HIE) follows; see APPENDIX D for citation conventions

13th) STATUS — Preliminary

Choose this entry since this is the first pass at the variable; its definition will probably be adjusted somewhat before it achieves semi-permanent status.

14th) VAR CREATN DATE — 10/10/85

This is the date the program ran that created the variable.

15th) ANALYST — Ron Hays

This is the analyst who created the variable; often a programmer's name is added afterwards.

16th) OBSERVER — Patient

The analyst is now the person whose view is being presented.

17th) SPACE — Universal

The respondent's income, which is the basis for this derived variable, will be the same regardless of where the respondent is

18th) TIME — 1985

The date of the raw variable is 1985 and there is nothing in the derived formula that alters that fact.

19th) FOOTNOTES — (1) SCORING: Log income (PPIPAT10):
log (minimum value + maximum value)/2, or
log 100000 INPUT VARIABLES: PPIPAT10
MISSING DATA RULE: No missing value imputations
(2) Var used to initially describe study patients
(3) One of three (PINC, PADJINC) income variables
(4) 10/15/86 Template reviewed by Ron Hays

(1) Variable construction description
(2) Analyst comments/reason for variable creation
(3) Other variables in the series
(4) Recency of template information

The resulting template:

VARIABLE ID:	PINCLOG
DATA LOCATION:	R.R5500.A4195.PDERIVED.SAS
PROG SOURCE:	/a/p/programs/derived5
PRIOR SOURCE:	MOS memo510, RAND HIE study
STATUS:	Preliminary

VAR CREATN DATE:	10/10/85
ANALYST:	Ron Hays
OBSERVER:	Patient
SUBJECT:	All screening patients
TOPIC:	Patient characteristics
SUBTOPIC:	Demographics/Socioeconomic status
MATTER:	Patient
FUNCTION:	Income
SPACE:	Universal
TIME:	1985
ASPECT:	Log of (F) Income
DOMAIN:	[1:infinity]
FOOTNOTES:	(1) SCORING: log income: log (lower limit + upper limit)/2, or log 100000 INPUT VARIABLES: PPIPAT10 MISSING DATA RULE: No imputation for missing values (2) Var used to initially describe study patients (3) One of three (PINC, PADJINC) income variables (4) 10/15/86 Template reviewed by Ron Hays

PART 3 - REPORTING AND MANIPULATING TEMPLATE INFORMATION

I. OVERVIEW

Resistence To Template Documentation

A critical question a documentor must ask him/herself after the templates have been so thoughtfully and laboriously created, is how useful the templates are. First experiences with MOS staff illustrated significant resistance to the templates. Even the most open-minded analyst called template terminology "voodoo mumbo-jumbo" and the number of facets grouped into one body was considered to make them difficult to read.

Overcoming Resistance

At Rand, we dealt with the first problem simply by teaching the analysts the terminology and meaning of the variables (since these are based on years of work by other individuals in the Language of Data group, we felt it our duty to preserve the labels they used for their descriptive elements), and renaming other less essential facets to fit MOS vernacular.

A Report Aa A Solution To Readability

As for the question of readability, we devised a report format that was much better suited to visual inspection. This benefited not only the primary analyst, but also satisfied criteria for comprehension of the public codebook.

A Report As A Means Of Augmenting Template Information With Information From Other Sources

Besides readability, the report allowed us to incorporate other sources of variable information into the codebook. For example, an analyst likes to see what raw variables comprise a derived variable. An easy way to produce such information is to copy a brief description of the variable from a data dictionary, rather than repetitiously retype the text. Another important source of information about the variable is statistics. This will be further discussed below.

II. Adding Dynamic Information (Statistics) To The Report

By now one would think that everything one could possibly want to know about the variables has been recorded. Not true. There is a category of dynamic information that we do not record in the template, but that study analysts need. Mainly, these are descriptive statistics. Because this information often changes as the dataset becomes more complete we do not update it regularly, but find it more practical to include it just before it will be viewed — in the reporting stage.

Reported Statistics

The statistics we include in our report are: frequencies — including percent, cumulative frequencies, cumulative percents, and means — including mean, standard deviation, minimum value, and maximum value.

Another element included in the statistics section is a reliability indicator. Reliability is a correlation of the true value of a variable and the analyst's measurement of the variable. If these two values are exactly the same (too idealistic to occur frequently) the correlation would be "1." A correlation of about .82 is more likely; most correlations between true and measured variables are between 0 and 1.

Contextual Information On Statistics

MOS analysts also found information regarding the creation of these figures most helpful so we included the date these statistics were created, the number of observations in the input dataset and the percentage of the dataset represented by missing observations. Information on the point at which an observation becomes missing (e.g., if 4 responses of the 8 input variables are missing, the derived variable is not calculated) is stored in the program creating the derived variable.

III. SAMPLE REPORT

VARIABLE ID: PCSTYLE

STUDY CONTEXT INFO

SUBJECT:	All Panel patients
TOPIC:	Style of Care
SUBTOPIC:	Interpersonal Style
SUB-SUBTOPIC:	Level of Patient Participation

VARIABLE MEANING

	OBSERVER: Patient
MATTER:	Patient/Clinician dyad
FUNCTION:	Clinician's willingness to share responsibility
SPACE:	Universal
TIME:	Screening Date (PSCRND (PATSCRND)
ASPECT:	Patient's Opinion of (F) Clinician's willingness to share responsibility
DOMAIN:	[0;100=Dyad shares responsibility]

VARIABLE CONSTRUCTION

	SCORING: Average of available items that have been recoded (where necessary) to fit direction of scale; final scale transformed to 0–100
INPUT VARIABLES:	SP1PST01 doctor CHOICE, ask to help make decision? SP1PST02 doctor INFOEX–answer questions politely?
COMMENTS:	1st in a series of 5 delineating...

STATISTICS

MEANS:	MEAN:	83.853	MINIM:	0.00	OBS:	10346
DATE:	STD:	15.855	MAXIM:	100.00	% MISS:	

FREQUENCIES:	VALUES:	FREQS	PERCNT:	CUM FREQ:	CUM PRCT:
DATE:	1	10	25	10	25
OBS:	2	10	25	20	50
MISSING:	3	10	25	30	75
	4	10	25	40	100

EXTERNAL INFORMATION

	DATA LOCATION: R.R0062.A4195.PDER2 SASNAME: PDERIVED
PROGRAM SOURCE:	/ma/p/programs/pdv.part2 (SP1 form only)
PRIOR SOURCE:	Memo714, memo664, memo644, memo643,
VARIABLE STATUS:	Semi-permanent
VAR CREATION DATE:	6/86
ANALYST:	Sherrie Kaplan, Ron Hays
ADMIN INDICATOR:	09/22/86 template reviewed by Ron Hays

IV. POSSIBILITIES FOR MANIPULATING TEMPLATE INFORMATION

We have been considering several options for implementing this system on a minicomputer or networked microcomputers. It is important that the computers be networked in order that information developed by one user can be easily shared by others.

There are several parts to the proposed system.

1. A display manager.

   A. This could be a text editor, possibly programmed to do interesting things with function keys. The pages would be formatted so that editor Page–Up and Page–Down functions would carry the user from one part of the documentation to another.

   B. A special–purpose viewer that could arrange and compare fields. The Language of Data's Descriptor Manipulator (Franzen, 1986) was an example of this.

   C. A multi–level display manager that worked through several levels: (a) the topic tree, (b) sorted lists of variables meeting given criteria, (c) an elaborated display of information about a given variable, and (d) a text editor for entering user comments.

2. A display generator.

   LOD experience suggests that the physical format of the documentation has a considerable effect on the ability of the user to find the correct information. Our data collection format, while efficient for data entry, does not particularly lend itself to casual viewing because there are too many elements and they all start on the left. Utility programs should be written to convert to more suitable viewing formats.

3. Acquisition of Statistics.

   Utility programs are required to gather the necessary statistics and insert them in appropriate spots. We anticipate unix scripts, or the equivalent, to do this.

4. Database Management.

   The documentation is itself a database, and needs to be accessed as such. Utility programs are required to convert from the input text format to database format. Retrieval operators are also needed for standard relational database purposes.

5. Program Generation

   A system of this kind should help assemble user programs to retrieve selected data from the database.

APPENDIX A

QUICK TEMPLATE REFERENCE

RAW VARIABLE TEMPLATE:

VARIABLE ID:	See /a/p/reference/instr.codes for conventions
DATA LOCATION:	where the dataset that contains the variable resides
PROG SOURCE:	where the program that creates the variable resides
PRIOR SOURCE:	MOS memo, bibliographic citation, another study
OBSERVER:	person answering questionnaire; respondent
OBSERVATION DATE:	(variable containing) date instrument fielding
SUBJECT:	Pool of people to whom the variable applies
TOPIC:	from MOS Conceptual Framework or MOS Concept Keys
SUBTOPIC:	from MOS Conceptual Framework or MOS Concept Keys
SUB-SUBTOPIC:	from MOS Conceptual Framework or MOS Concept Keys
MATTER:	object under discussion (patient, clinician, visit
FUNCTION:	event happening to matter or matter's state-of-being
SPACE:	where event took place; For N/A use: Universal
TIME:	time the var refers to (1984 income, hlth last week)
ASPECT:	specific character of matter of function
DOMAIN:	possible value range: ex: yes/no, 1-4, etc
FOOTNOTES:	(1) Question text
	(2) Analyst Comments/experiences re: variable
	(3) Series information here
	(4) Template indicator noting recency of information

DERIVED VARIABLE TEMPLATE:

VARIABLE ID:	See /a/p/reference/derived.suffix for conventions
DATA LOCATION:	where data resides on the computer
PROG SOURCE:	source [programs] for data creation
PRIOR SOURCE:	MOS memo, bibliographic citation, another study
STATUS:	PROPOSED, PRELIMIARY, SEMI-PERMANENT OR PERMANENT;
VAR CREATN DATE:	date of dataset creation or MOS memo defining var
ANALYST:	Main analyst(s), programmer
OBSERVER:	specify the original observer's name here
SUBJECT:	Pool of people to whom the variable applies
TOPIC:	from MOS Conceptual Framework or MOS Concept Keys
SUBTOPIC:	from MOS Conceptual Framework or MOS Concept Keys
MATTER:	object under discussion (patient, clinician, visit)
FUNCTION:	event happening to matter or matter's state-of-being
SPACE:	where event took place; For N/A use: Universal
TIME:	time the var refers to (1984 income, hlth last week)
ASPECT:	specify character of function or matter

DOMAIN: possible value range; ex: yes/no, 1-4, etc.

FOOTNOTES: (1) Variab!e construction (SCORING, INPUT VARS, MISSING DATA RULE)
(2) Analyst Comments (assumptions, experiences, reason for creating the variable)
(3) Series information here
(4) Template indicator noting recency of information

APPENDIX B

A RAW VARIABLE TEMPLATE VS. A DERIVED VARIABLE TEMPLATE

VARIABLE ID:	SP4INC10
DATA LOCATION:	/a/p/datasets/SP4
SOURCE:	MOS memo315,
PRIOR SOURCE:	RAND HIE study
OBSERVATION DATE:	Screener week10/10/85
ANALYST:	Ron Hays
OBSERVER:	Patient
SUBJECT:	All screening patients
TOPIC:	Patient characteristics
SUBTOPIC:	Demographics/Socioeconomic status
MATTER:	Patient
FUNCTION:	Income
SPACE:	Universal
TIME:	1985
ASPECT:	Amount of (F) Income
DOMAIN:	1 = $0 to $4999,
	2 = $5000 to $9999,
	3 = $10000 to $14999,
	4 = $15000 to $19999,
	5 = $20000 to $29999,
	6 = $30000 to $49999,
	7 = $50000 to $99999,
	8 = $100000 and up
FOOTNOTES:	(1) What was your total household income before taxes in 1985?
	(2) Income categories are used for the response set instead of asking a specific amount since some people consider the latter an invasion of privacy
	(4) 10/15/86 Template reviewed by Ron Hays

VARIABLE ID:	PINCLOG
LOCATION:	R.R5500.A4195.PDERIVED.SAS
SOURCE:	/a/p/programs/derived5
PRIOR SOURCE:	MOS memo no. 315, RAND HIE study
STATUS:	Prelimiary
VAR CREATN DATE:	10/10/85
ANALYST:	Anita Steward
OBSERVER:	Patient
SUBJECT:	All screening patients
TOPIC:	Patient characteristics
SUBTOPIC:	Demographics
MATTER:	Patient
FUNCTION:	Income

SPACE:	Universal
TIME:	1985
ASPECT:	Log of (F) Income
DOMAIN:	[1:infinity]
FOOTNOTES:	(1) SCORING: log income: log (lower limit + upper limit)/2, or log 10000 INPUT VARIABLES: PPIPAT10 MISSING DATA RULE: No imputation for missing values (2) Var used to initially describe study patients (3) One of three (PINC, PADJINC) income variables (4) 10/15/86 Template reviewed by Ron Hays

APPENDIX C

DESCRIPTIVE ELEMENTS ENTRY PRIORITIES

It is helpful to keep the following issues in the following order:

- the aspect must originate from the domain
the aspect must refer back to either matter, function, or time
function must directly connect with with subtopic (or sub–subtopic)
within the MOS, subtopic must fit in with the overview concepts list

Overview, Conceptual Framework and Key study variables

Below is an example of what can happen when the above is not considered and/or whether the analyst is consulted about the intent of the variable.

e	Initial	Classification
e	Classification	following order
VARIABLE ID:	SP2HYPO5	SP2HYPO5
DATA LOCATION:	S.S8587.A4195.SP2SAS	S.S8587.A4195.SP2SAS
SOURCE:	/a/p/programs/pdv.part1	/a/p/programs/pdv.part1
PRIOR SOURCE:	MOS Memo here	MOS Memo here
OBSERVER:	Screener week	Screener week
OBSERVATION DATE:	Patient	Patient
SUBJECT:	1/2 of screening patients	1/2 of screening patients
	*******************	***************************
TOPIC:	* Process of Care *	* Patient characteristics *
SUBTOPIC:	* Treatments *	* Disease Severity *
	*******************	***************************
SUB-SUBTOPIC:	(None)	Hypertension
MATTER:	Patient	Patient
	****************	**************
FUNCTION:	* Hypertension *	* Drug Usage *
	****************	**************
SPACE:	Universal	Universal
TIME:	Current	Current
	***************************	****************************
ASPECT:	* Treatment of (F) Hyper- *	* Number of (F) Drugs taken *
	* tension *	* for Hypertension *
	***************************	****************************
DOMAIN:	1=None, 2=One, 3=Two or more	1=None, 2=One, 3=Two or more

FOOTNOTES:	(1) How many different drugs are you taking for your high blood pressure?	(1) How many different drugs are you taking for your high blood pressure? (3) Assumes that worse cases require more drugs to control hypertension.

In the initial classification, because of the text of the question, the archivist assumed that SP2HYPO2 is about Treatment and Process of Care. However, after verifying with the analyst, the archivist found that SP2HYPO2 is intended to measure patient's disease severity.

APPENDIX D

BIBLIOGRAPHIC CITATIONS

Rand standard bibliographic citations are and memo citations should be in the following form:

MEMOS	REPORTS	JOURNALS	BOOKS
author	author	author	author
title quotes	title quotes	title quotes	title quotes
memo num	report num	volume num	
	publisher/company	publisher/company	publisher
	place	place	place
date	date	date	date
page numbers	page numbers	page numbers	page numbers

In keeping with general syntax, each citation should be separated by commas. Therefore, any commas e.g., ones separating author, title, etc. in the citation should be replaced by semi-colons.

Another type of reference is for datasets, e.g. AMA data. A format could be:

author company
title and date yr of data collection
Rand Data facility database number
publisher/distributor
date data was published

REFERENCES

Dolby, J.L.; Clark, N., 1982: *The Language of Data.* San Jose State University Foundation.

Dolby, J.L., 1983: Meaning from Data, *Proceedings of the AAAS*, May 1983.

Dolby, J.L.; Clark, N.; Rogers, W.H., 1986: The Language of Data: A General Theory of Data, *Proceedings, Eighteenth Symposium on the Interface*, American Statistical Association.

Clark, N., 1985: *Classification Procedures: Classification from Survey Instruments.* Technical Report, San Jose State University Department of Mathematics and Computer Science. For copies contact Nancy Clark, Box R, Sausalito, CA 94966.

Clark, N., 1987: Tables and Graphs as a Form of Exposition. *Scholarly Publishing.* October 1987, in press.

Ware, J. et al, 1985: *A National Study of Medical Care Outcomes: Project Overview.* Contact John Ware, The Rand Corporation, Santa Monica, CA 90406 for copies.

New Data Sets From Rand

NEW RESEARCH DATA FILES AND DOCUMENTATION

The **RAND** Corporation is releasing files of research data gathered in the Health Insurance Experiment (HIE), a major project conducted by RAND from 1974 to 1982. The experiment was a large-scale, controlled trial in health care financing, funded by the U.S. Department of Health and Human Services. Its purpose was to assess the effects of different types of insurance on patient health and health care delivery, including both fee-for-service (FFS) and health maintenance organization (HMO) modes.

Over 8,200 persons were enrolled in the experiment in six sites: Dayton, Ohio; Seattle, Washington; Franklin County and Fitchburg, Massachusetts; and Georgetown County and Charleston, South Carolina. Each family was assigned to one of a number of insurance plans that varied by coinsurance rate, delivery mode, and maximum out-of-pocket expenditure. Data were collected on enrollees' use of health care services and state of health throughout their term of enrollment, three or five years. A large subset of the collected data is being released in 67 fully documented files.

Documentation for each file includes a data tape and hardcopy codebook published as a RAND Note.

* The machine-readable tape contains three datasets: data in Statistical Analysis System (SAS) format, data in character format, and a dictionary for the character-format file.

* The codebook provides an overview of the experiment, outlines the characteristics of the file and related files, and describes each variable. Variable descriptions include response codes and response frequencies or summary statistics.

Codebooks can be ordered from the RAND Publications Department; indicate the desired titles and N numbers. Data files can be ordered from the RAND Data Facility. For each file desired, indicate the file reference number.

"MASTER SAMPLE SERIES"

MS1: Eligibility-Family Changes File. Data on insurance status, changes in eligibility status, and family relationships. No other file provides this information.

Sample: Insured enrollees (including members of HMO control group) and persons of interest, a subset of adjunct enrollees. Total of 9,142 records, one per person.

Representative Variables (of the total of 54): Insurance status, start/end dates of insurance coverage, reason for losing coverage, length of time insured, family identifier, individual's relation to family/household head.

Codebook: N-2264/1-HHS, Vol. 1: *Codebook for Eligibility-Family Changes File* , by S. M. Polich and C. d'Arc Taylor, May 1986.

MS2: Full Sample Demographic File. Baseline, enrollment, and demographic data. No other file provides this information.

Sample: All HIE participants—insured enrollees, adjunct enrollees, and baseline-only participants. Total of 26,148 records, one per person.

Representative Variables (total, 62): Date of baseline interview, sex, age, race, marital status, education, income, occupation, health insurance, welfare status, hospitalizations, family doctor, medical/dental visits and expenses, assigned experimental insurance plan.

Codebook: N-2264/2-HHS, Vol. 2: *Codebook for Full Sample Demographic File,* by S. M. Polich et al., May 1986.

MS3: Supplemental Data File. Sample data needed for specific analyses. No other file provides this information.

Sample: Overall sample is the same as that in the full sample demographic file; specific sample differs by variable. Total of 26,148 records, one per person.

Representative Variables (total, 19): Changes affecting HMO and FFS-HMO analyses, enrollment refusals, identifier for mothers of newborns, revised death date.

Codebook: N-2264/3-HHS,¯Vol. 3: *Codebook for Supplemental Data File,* by S. M. Polich and C. d'Arc Taylor, June 1987.

"CLAIMS LINE-ITEM SERIES"

LI1-LI14: FFS Claims Line-Item Files (14 files). Detailed FFS claims data, including data for HMO enrollees who used medical/dental services in FFS sector. Number of variables in each file shown parenthetically below.

LI1. Hospital inpatient services (39)
LI2. Inpatient physician procedures billed by institutions (36)
LI3. Drugs prescribed by physicians (52)
LI4. Supplies prescribed by physicians (44)
LI5. Services rendered by physicians (53)
LI6. Drugs sold by physicians (63)
LI7. Supplies sold by physicians (53)
LI8. Injections administered by physicians (69)
LI9. Outpatient services billed by institutions (45)
LI10. Services rendered by dentists (33)
LI11. Drugs prescribed by dentists (24)
LI12. Drugs purchased (37)
LI13. Supplies purchased from pharmacies (17)
LI14. Supplies purchased from nonpharmacy suppliers (19)

Sample: Insured enrollees who filed claims. Total of 603,998 records, one per line item.

Representative Variables: Diagnoses (multiple), provider ID, inpatient/outpatient procedures/services, drugs prescribed and sold, dosage instructions, symptoms, relation to employment/accident, treatment history, charges, supplies bought.

Codebook: N-2347/1-HHS, Vol. 1: *Codebooks for Fee-for-Service Claims*, by C. E. Peterson et al., June 1986.

LI15-LI25: HMO Claims Line-Item Files (11 files). Detailed services provided or reimbursed by the HMO. Number of variables in each file shown parenthetically below.

LI15. Hospital inpatient services (34)
LI16. Inpatient physician services (41)
LI17. Drugs prescribed by physicians (52)
LI18. Supplies prescribed by physicians (45)
LI19. Services rendered by physicians (51)
LI20. Drugs dispensed by physicians (57)
LI21. Supplies dispensed by physicians (44)
LI22. Injections administered by physicians (66)
LI23. Outpatient services provided by institutions (42)
LI24. Drugs dispensed (32)
LI25. Supplies dispensed (13)

Sample: HMO participants, a subset of insured Seattle enrollees. Total of 177,566 records, one per line item.

Representative Variables: Diagnoses, provider ID, procedures performed, drugs/supplies, imputed charges.

Codebook: N-2347/2-HHS, *Vol. 2: Codebooks for Health Maintenance Organization Claims*, by C. E. Peterson et al., August 1986.

LI26-LI29: FFS Claims for FFS-HMO Comparison (4 files). Detailed Seattle FFS claims data with imputed charges for physician services to enable dollar comparisons with HMO data. Number of variables in each file shown parenthetically below.

LI26. Hospital inpatient services (35)
LI27. Inpatient physician procedures billed by institutions (32)
LI28. Outpatient services rendered by physicians (50)
LI29. Injections administered by physicians (64)

Sample: Insured Seattle FFS enrollees who filed claims. Total of 70,991 records, one per line item.

Representative Variables: Diagnoses, provider ID, imputed charges, procedures performed.

Codebook: N-2347/3-HHS, Vol. 3: *Codebooks for Seattle Fee-for-Service Claims for Comparison with Health Maintenance Claims*, by C. E. Peterson, M. Nelsen, and D. L. Wesley, October 1986.

"AGGREGATED CLAIMS SERIES"

AC1: FFS Annual Expenditure File. Claims data aggregated by year for insured FFS enrollees. Also covers FFS dental usage by HMO enrollees.

Sample: Insured enrollees. Total of 25,740 records, one per enrollee per year.

Representative Variables (total, 23): Number per year of the following: hospitalizations, physician and nonphysician visits, mental health visits, and dental visits; annual expenditures for inpatient, outpatient, mental health, and dental services.

Codebook: N-2360/1-HHS, Vol. 1: *Codebook for Fee-for-Service Annual Expenditures and Visit Counts,* by C. E. Peterson, M. Nelsen, and E. S. Bloomfield, May 1986.

AC2-AC4: FFS Visit Files (3 files). Claims data aggregated by outpatient, inpatient, and dental visit for insured FFS enrollees. Covers
FFS dental visits by HMO enrollees. Number of variables in each file shown parenthetically below.

AC2. FFS outpatient visits (46)
AC3. FFS inpatient visits (55)
AC4. FFS dental visits (16)

Sample: Insured enrollees. Total of 148,123 records, one per enrollee-provider-date of service.

Representative Variables: Type of visit, providers, visit dates, procedures, diagnoses, charges.

Codebook: N-2360/2-HHS, Vol. 2: Codebooks for Fee-for-Service Visits—Outpatient, Inpatient, and Dental, by C. E. Peterson et al., June1986.

AC5-AC6: FFS Episode Files (2 files). Claims data aggregated by treatment episode for FFS enrollees. Description and expenses for each episode (individual file); episode counts and expenses per year (annual file). Number of variables in each file shown parenthetically below.

AC5. FFS individual episodes (19)
AC6. FFS annual episodes (42)

Sample: Insured FFS enrollees. Total number of records: 99,001 in individual file, one per episode; 21,094 in annual file, one per enrollee per year.

Representative Variables: Episode description, start/end dates, diagnosis, expense limit at beginning of year, remaining expense limit at start/end of episode, number of episodes per year by type, expenses per episode type per year.

Codebook: N-2360/3-HHS, Vol. 3: *Codebooks for Fee-for-Service Treatment Episodes and Annual Episode Counts,* by C. E. Peterson, C. d'Arc Taylor, and E. S. Bloomfield, June 1986.

AC7: HMO and Seattle FFS Annual Expenditure File. Claims data aggregated by year for insured HMO and Seattle FFS enrollees.

Sample: Insured Seattle enrollees. Total of 11,221 records, one per enrollee per year.

Representative Variables (total, 33): Number per year of the following: hospitalizations, physician and nonphysician visits, mental health visits, imputed expenditures

Codebook: N-2360/5-HHS, Vol. 5: *Codebook for Health Maintenance Organization and Seattle Fee-for-Service Annual Expenditures and Visit Counts,* by C. E. Peterson et al., December 1986.

AC8-AC9: HMO and Seattle FFS Visit Files (2 files). Claims data aggregated by health care visit for insured HMO and Seattle FFS enrollees. Number of variables in each file shown parenthetically below.

AC8. HMO and Seattle FFS outpatient visits (45)
AC9. HMO and Seattle FFS inpatient visits (53)

Sample: Insured Seattle enrollees. Total of 61,597 records, one per enrollee-provider-date of service.

Representative Variables: Type of visit, providers, visit dates, procedures, diagnoses, imputed charges.

Codebook: N-2360/4-HHS, Vol. 4: Codebooks for Health Maintenance Organization and Seattle Fee-for-Service Visits—Outpatient and Inpatient, by C. E. Peterson, M. Nelsen, and D. L. Wesley, December 1986.

"HIE REFERENCE SERIES"

RF1: Codes (no tape file). N-2349/1-HHS, Vol. 1: Codes Used in HIE Claims—Diagnoses, Symptoms, Procedures, Drugs, and Supplies, by M. Nelsen and C. A. Edwards, May 1986. Defines all codes used in HIE claims data, both line-item and aggregated. Includes standard and HIE-created codes: diagnosis (H-ICDA-2), CRVS, supply, reason for visit/symptom, NDC, generic drug, drug therapeutic, American Dental Association procedure. **

**-ICDA-2 refers to the second version of the hospital adaptation of International Classification of Diseases Adapted for Use in the United States; CRVS refers to codes for medical and surgical procedures taken from California Relative Value Studies; NDC refers to National Drug Code.

RF2: HIE Provider File: Information about the physicians, hospitals, dentists, and other providers of services to HIE enrollees.

Sample: All providers cited in HIE data. Total of 22,658 records, one per provider identifier.

Representative Variables (total, 26): Provider type, provider specialty, linking identifier.

Codebook: N-2349/2-HHS, Vol. 2: *Providers Cited in HIE Data,* by S. M. Polich, M. Nelsen, and D. L. Wesley, June 1987.

RF3: User's Guide (no tape file). N-2349/3-HHS, Vol. 3: User's Guide to HIE Data, by C. d'Arc Taylor, S. M. Polich, C. E. Peterson, and E. M. Sloss, August 1987. Analytic possibilities and limitations of HIE data; suggestions for choosing analytic subsamples and linking data across series, files, years, sample groups, and sites for particular analytic purposes.

"MEDICAL HISTORY QUESTIONNAIRE SERIES"

MH1A-MH3A: Adult Form A (3 files). Data from self-administered questionnaire on health status, attitudes, and habits. Number of variables in each file shown parenthetically below.

MH1A. Dayton adults at enrollment, form A (364)
MH2A. NonDayton adults at enrollment, form A (373)
MH3A. Adults at exit, form A (408)

Sample: Adults (14 and older) when enrolling and when completing assigned term three or five years later. Includes insured enrollees, Dayton control group, and PEG-period-only participants. Total of 9,058 records, one per completed questionnaire.

Topics: height and weight, general health, eating habits, sleep and exercise, seat belt use, smoking and drinking, general well being, social activities, life events, symptoms, health perceptions, medical opinions, medical and dental care, effectiveness of health care.

Codebook: N-2485/1-HHS, *Vol. 1: Codebooks for Adults at Enrollment and Exit, Form A,* by C. A. Edwards et al., August 1986.

MH1B-MH3B: Adult Form B (3 files). Data from self-administered questionnaire on verifiable physical limitations and specific medical disorders. Number of variables in each file shown parenthetically below.

MH1B. Dayton adults at enrollment, form B (282)
MH2B. NonDayton adults at enrollment, form B (480)
MH3B. Adults at exit, form B (490)

Sample: Adults (14 and older) when enrolling and when completing assigned term three or five years later. Includes insured enrollees, Dayton control group, and PEG-period-only participants. Total of 9,914 records, one per completed questionnaire.

Topics: Vision, hearing, hay fever, teeth and gums, fluoride treatment, acne, thyroid, joints, heart/lung ailments, hypertension, stroke, stomach, kidney/bladder, cholesterol, anemia, diabetes, cancer, surgical conditions (hemorrhoids, hernia, varicose veins), physical/activity limitations, sleeping pill use, missing limbs, antibiotic allergy, effectiveness of health care, immunization, gall bladder/tonsil surgery, female organs, medical care, medical appliances, future health expenses, transportation for health care.

Codebook: N-2485/2-HHS, Vol. 2: *Codebooks for Adults at Enrollment and Exit, Form B,* by C. A. Edwards et al., October 1986.

MH4A-MH6B: Child Forms A and B (6 files). Data from two parent-completed questionnaires.

Form A pertained to health status, attitudes, and habits. Form B pertained to verifiable physical limitations and specific medical disorders. Number of variables in each file shown parenthetically below.

MH4A. Dayton children at enrollment, form A (85)
MH4B. Dayton children at enrollment, form B (63)
MH5A. NonDayton children at enrollment, form A (151)
MH5B. NonDayton children at enrollment, form B (224)
MH6A. Children at exit, form A (147)
MH6B. Children at exit, form B (235)

Sample: Children 5-13 years old, at family's enrollment and exit three or five years later. Includes insured enrollees, Dayton control group, and PEG-period-only participants. Total of 6,958 records, one per completed questionnaire.

Topics: Form A: height, weight, general health, fluorides, diet, immunizations, safety practices, learning, getting along, general well-being, symptoms, behavior problems. Form B: teeth, fluoride treatment, eyesight, hearing, ear infections, asthma, hay fever, eczema, anemia, lead poisoning, kidney/bladder infection, bedwetting, cancer, convulsions, tonsils, antibiotic allergy, missing limbs, medical appliances, future health expenses.

Codebook: N-2485/3-HHS, Vol. 3: *Codebooks for Children at Enrollment and Exit,* by C. A. Edwards et al., November 1986.

MH7A-MH9B: Infant Forms A and B (6 files). Data from two parent-completed questionnaires. Form A pertained to health status and development; form B pertained to verifiable physical limitations and specific medical disorders. Number of variables in each file shown parenthetically below.

MH7A. Dayton infants at enrollment, form A (76)
MH7B. Dayton infants at enrollment, form B (28)
MH8A. NonDayton infants at enrollment, form A (98)
MH8B. NonDayton infants at enrollment, form B (122)
MH9A. Infants at exit, form A (94)
MH9B. Infants at exit, form B (134)

Sample: Infants (0-4 years old) at family's enrollment and exit three or five years later. Includes insured enrollees, Dayton control group, and PEG-period-only participants. Total of 3,334 records, one per completed questionnaire.

Topics: Form A: height, weight, development, general health, fluorides, diet, immunizations, safety practices, symptoms. Form B: colds, ear infections, eczema, anemia, lead poisoning, cancer, convulsions, tonsils, antibiotic allergy, missing limbs, medical appliances, future health expenses, fluoride treatment.

Codebook: N-2485/4-HHS, Vol. 4: *Codebooks for Infants at Enrollment and Exit,* by C. A. Edwards et al., December 1986.

"HEALTH STATUS AND ATTITUDE SERIES"

HS1-HS2: Adult and Child (2 files). Data derived from medical history questionnaire on enrollees' state of health and attitudes toward health care at enrollment and exit. Number of variables in each file shown parenthetically below.

HS1. Adults at enrollment and exit (136)
HS2. Children at enrollment and exit (28)

Sample: Insured enrollees: adults (14 and older) and children (aged 0-13) at family's enrollment and exit three or five years later. Total of 5,871 records in adult file, 2,840 records in child file—one per person per file.

Representative Variables: Scales of physical health, mental health, social health, perceptions of general health; satisfaction with medical and dental care, cigarette smoking, alcohol consumption, weight, and exercise (adults only).

Codebook: N-2447/1-HHS, Vol. 1: *Codebooks for Adults and Children at Enrollment and Exit*, by E. M. Sloss et al., November 1986.

"MEDICAL DISORDERS SERIES"

MD1: Adults. Data derived from medical history questionnaire and medical screening examination on 17 disorders: acne, anemia, angina pectoris, chronic obstructive airway disease, congestive heart failure, diabetes mellitus, hay fever, hearing loss, hypercholesterolemia, hypertension, joint disorders, kidney disease and urinary tract infection, peptic ulcer disease, sleeping pill and tranquilizer use, surgical conditions, thyroid disease, and vision disorders.

Sample: Insured enrollees: adults (14 and older) at enrollment and exit three or five years later. Total of 5,871 records, one per person.

Topics (total number of variables, 286): Status and severity of disorder, impact of disorder, results of medical tests.

Codebook: N-2446/1-HHS, *Vol. 1: Codebook for Adults at Enrollment and Exit*, by B. H. Operskalski et al., February 1987.

MD2: Children. Data derived from medical history questionnaire and medical screening examination on four disorders: allergic conditions, anemia, middle ear disease and hearing impairment, and vision impairment.

Sample: Insured enrollees: children (aged 0-13) at family's enrollment and exit three or five years later. Total of 2,840 records, one per person.

Topics (total number of variables, 73): Status and severity of disorder, impact of disorder, results of medical tests.

Codebook: N-2446/2-HHS, Vol. 2: *Codebook for Children at Enrollment and Exit*, by E. M. Sloss et al., March 1987.

"DENTAL EXAMINATIONS FILE"

DE1: Dental Examinations File. Data from a dental screening examination on tooth decay and its consequences, and periodontal disease and its severity.

Sample: Insured enrollees: persons aged three and older in randomly selected subsample (50-75 percent) of families at enrollment; all persons aged three and older at exit three or five years later. Total of 7,317 records, one per person.

Representative Variables (total, 50): Number of decayed primary and permanent teeth, number of missing or extracted primary and permanent teeth, number of filled primary and permanent teeth, oral hygiene index score, and (for those 12 and older) periodontal disease index score.

Codebook: N-2506-HHS, Dental Examinations: *Codebook for Adults and Children at Enrollment and Exit*, by E. S. Bloomfield, L. Y. Weissler, and A. M. Bell, February 1987.

"INSURANCE PREFERENCE FILES"

IP1-IP2: Insurance Preference Files (2 files). Data from self-administered questionnaire on willingness to pay a higher health insurance premium in return for a lower annual out-of-pocket expense limit—three hypothetical premium-expense limit combinations. Number of variables in each file shown parenthetically below.

IP1. Maximum-dollar-expenditure plans (21)
IP2. Fixed-dollar-limit plan (25)

Sample: Heads of insured enrollee families (except HMO enrollees and enrollees assigned to receive free care) when completing assigned enrollment term. Total of 2,020 records, one per questionnaire recipient.

Topics: Family's expense limit during its last year in HIE; premium-expense limit combination for each of three hypothetical offers; degree of willingness to accept each offer.

Codebook: N-2508-HHS, *Codebooks for Insurance Preference Files: Relation between Expense Limit and Premium*, by E. S. Bloomfield, L. Y. Weissler, and A. B. Holland, October 1986.

Order form

Place a check by the HIE data tapes or publications you wish to order. If you want more than one copy, indicate the number desired. Send the completed form, along with your address, to CIS Business Office, The RAND Corporation, P.O. Box 2138, Santa Monica, CA 90406-2138.

Data Tapes*			Publications
Master Sample Series			
__MS1	$ 50.00	__N-2264/1-HHS	$ 7.50
__MS2	50.00	__N-2264/2-HHS	15.00
__MS3	50.00	__N-2264/3-HHS	7.50
Claims Line-Item Series			
__LI1-LI14	150.00	__N-2347/1-HHS	25 00
__LI1-LI14			
(SAS version only)	100.00		
__LI15-LI25	100.00	__N-2347/2-HHS	25.00
__LI26-29	50.00	__N-2347/3-HHS	15.00
Aggregated Claims Series			
__AC1	50.00	__N-2360/1-HHS	7.50
__AC2-AC4	100.00	__N-2360/2-HHS	15.00
__AC5-AC6	100.00	__N-2360/3-HHS	10.00
__AC7	50.00	__N-2360/5-HHS	10.00
__AC8-AC9	100.00	__N-2360/4-HHS	15.00
HIE Reference Series			
__RF1 (no tape exists)		__N-2349/1-HHS	20.00
__RF2	50.00	__N-2349/2-HHS	7.50
__RF3 (no tape exists)		__N-2349/3-HHS	10.00
Medical History Questionnaire Series			
__MH1A-MH3A	50.00	__N-2485/1-HHS	25.00
__MH1B-MH3B	50.00	__N-2485/2-HHS	25 00
__MH4A-MH6B	50.00	__N-2485/3-HHS	25.00
__MH7A-MH9B	50.00	__N-2485/4-HHS	20.00
Health Status and Attitude Series			
__HS1-HS2	50.00	__N-2447/1-HHS	25 00
Medical Disorders Series			
__MD1	50.00	__N-2446/1-HHS	25.00
__MD2	50.00	__N-2446/2-HHS	15 00
Other Files			
__IP1-IP250.00 __N-2508-HHS		7.50	
__DE1	50.00	__N-2506-HHS	7.50
__All 67 files	1000.00		

*Except where noted, each tape contains three datasets: (1) data in Statistical Analysis System (SAS) format, (2) data in character format, and (3) dictionary for character-format data.

DATASET NEWS

Available on Diskette

SURVEYS FOR THE AMERICANS TALK SECURITY PROJECT

The Americans Talk Security (ATS) Project is a series of nationwide surveys on American attitudes toward national security issues to be conducted during the year preceding the 1988 presidential election.The Project has funded a bipartisan group of four prominent opinion research organizations to cooperatively conduct a total of twelve surveys. Those doing the polling are:

Market Opinion Research, Marttila & Kiley, The Daniel Yankelovich Group, and the Public Agenda Foundation.

On a rotating schedule, each of the polling firms assumes primary responsibility for an individual survey. Every study is, however, reviewed and approved for fairness, balance, and accuracy by the other participating organizations. An ideological balance is maintained as two of the polling firms, the Daniel Yankelovich Group and the Public Agenda Foundation endeavor to remain politically neutral. Of the others, Marttila & Kiley do polling exclusively for Democratic candidates and Market Opinion Research exclusively for Republican candidates.

The ATS project works closely with the Democratic and Republican presidential campaigns, briefing them on the results of the surveys and offering to design and include in upcoming surveys questions of particular interest to them. In some instances this offer has brought forth questions not normally, and some never before, addressed in national surveys.

It is a key purpose of the project that the public be fully and widely informed of all results, and that the data see maximum utilization in the corporate and academic research communities. To that end ATS, in special cooperation with the Roper Center, is offering the datasets for each of the surveys on diskette (3 1/2 or 5 1/4 inch) at a cost of $35.00 per study. And ISLA members may charge the cost of the diskettes against their membership balance.

The ATS Project is processing all requests for these data. Should you wish to place an order you may contact them at:

Americans Talk Security
83 Church Street
Unit17
Winchester, MA 01890
(617) 721-9266

The Roper Center is pleased to assist the Americans Talk Security Project in this significant undertaking. We welcome your inquiries.

IASSIST

The International Association for Social Science Information Services and Technology (**IASSIST**) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and /or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST**.

Membership fees are:
Regular Membership. $20.00 per calendar year.
Student Membership: $10.00 per calendar year.

Institutional subcriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subcription: $35.00 per calendar year (includes one volume of the Quarterly)

Membership form

I would like to become a member of IASSIST. Please see my choice below:

$20 Regular Membership
$10 Student Membership
$35 Institutional Membership

My primary Interests are:

Archive Services/ Administration
Data Processing/Data Management
Research Applications
Other (specify) ________

Please make checks payable to IASSIST and Mail to :

Ms Jackie McGee
Treasurer, IASSIST
% Rand Corporation
1700 Main Street
Santa Monica

Name /phone ________

Institutional Affiliation ________

Mailing Address ________

City ________

Country /zip/postal code ________